Design adaptation methods for genetic association studies

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Humanbiologie
(Dr. rer. physiol.)
dem Fachbereich Medizin
der Philipps-Universität Marburg
vorgelegt von

ANDRÉ SCHERAG
aus Bad Hersfeld

Marburg, 2008
Angenommen vom Fachbereich Medizin der Philipps-Universität Marburg am 27.02.2008

Gedruckt mit Genehmigung des Fachbereichs.

Dekan: Prof. Dr. M. Rothmund

Referent: Prof. Dr. H. Schäfer

1. Korreferent: Prof. Dr. K.-H. Grzeschik
Contents

1 Introduction ............................................. 9

2 Genetic mapping of complex traits .................. 13
  2.1 Biological background .................................. 13
    2.1.1 Genetic linkage analysis ................................. 15
    2.1.2 Linkage disequilibrium mapping - genetic association analysis .... 16
  2.2 Statistical methods for genetic association analysis ...... 20

3 Mathematical background ................................. 25
  3.1 Adaptive designs and adaptive design methods ........... 26
    3.1.1 General framework ........................................ 26
    3.1.2 Adaptive designs ........................................... 27
    3.1.3 Methods for design adaptation ............................ 34
  3.2 Point and interval estimates of genetic effects ............ 36
    3.2.1 Case-control design ....................................... 36
    3.2.2 Family-based trio design ................................ 37

4 Sample size modifications for candidate gene studies .... 49
  4.1 Introduction ............................................ 49
  4.2 Method ................................................. 50
    4.2.1 Fixed sample, group sequential and data adaptive plans ........... 50
    4.2.2 Approximation of the TDT by a Brownian motion process - group sequential TDT .... 51
    4.2.3 Interim analysis and sample size adaptation .................... 53
Contents

4.3 Simulation studies .............................................. 55
  4.3.1 Comparison of FS, GS, and AS designs ................. 55
  4.3.2 Comparison with adaptive designs ....................... 56
  4.3.3 Modifying the rule for interim sample size reassessments .. 62
4.4 Real data application ........................................ 65
4.5 Discussion .................................................. 66

5 Flexible two-stage designs for genomewide association studies 71
  5.1 Introduction .................................................. 71
  5.2 Method ...................................................... 73
    5.2.1 Notation for a case-control genomewide association study ... 73
    5.2.2 The flexible two-stage procedure ....................... 75
  5.3 Simulation studies ......................................... 77
  5.4 Real data application ..................................... 82
  5.5 Discussion ................................................. 83

6 General discussion ........................................... 87
  6.1 Design adaptation methods for genomewide association studies .... 88
  6.2 Design adaptation methods for candidate gene investigations .... 90
  6.3 Limits and prospects of gene characterization using estimates of genetic
      effects ...................................................... 92

7 Summary ..................................................... 95

8 Zusammenfassung .............................................. 97

References ..................................................... 99

Lebenslauf ..................................................... 114

Verzeichnis akademischer Lehrer ................................. 118

Danksagung ..................................................... 119
Contents

Ehrenwörtliche Erklärung 121

Appendix 122
Summary of commonly used notation and abbreviations

Variables

\(\alpha, \gamma\)  
significance levels of a statistical test

\(c_{\alpha}\)  
critical value of a level \(\alpha\) test

\(F_{k_1,k_2,(1-\alpha)}\) denotes the \((1-\alpha)\)-quantile of the central \(F\)-distribution with \((k_1,k_2)\) df  
(in chapter 5, \(F\) is used to denote a cumulative distribution function)

\(i, j, k\)  
index variables

\(K, U, X\)  
random variables  
(bold letters for vectors, small letters for their realizations)

\(n, N\)  
variable and random variable for sample size

\(p\)  
p-value  
(not to be confused with the allele frequency abbreviation)

\(t\)  
information time variable  
(in chapter 5, \(t^{1-\alpha}\) is the \((1-\alpha)\)-quantile of the cumulative distribution function of a test statistic)

\(T\)  
test statistic

\(Z\)  
\(\mathcal{N}(0, 1)\)-distributed random variable

\(z_{1-\alpha}\)  
\((1-\alpha)\)-quantile of a standard normal distribution

\(\chi^2_{k,1-\alpha}\)  
\((1-\alpha)\)-quantile of a chi-squared distribution with \(k\) df
Parameters

$\Delta$  
LD measure

$\delta, \sigma^2$  
Brownian motion process drift and variance

$f = (f_0, f_1, f_2)$  
penetrance vector

$\pi$  
proportion of sample size

$\phi$  
proportion of cases relative to total sample size

$\Psi_1, \Psi_2$  
genotype relative risk parameters

$p, q$  
allele frequency of D and M

$\theta$  
recombination fraction

$\vartheta$  
parameter of a random distribution

$w = (w_0, w_1, w_2)$  
weights for CA Trend Test

Distributions, Measures, Functions, Processes, and sets

$\mathcal{N}(0, 1)$  
standard normal distribution

$\mathcal{U}(0, 1)$  
uniform distribution

$\alpha(\cdot)$  
conditional (type I) error function

$E(\cdot)$  
extected value

$\varphi(\cdot), \psi(\cdot)$  
decision functions

$I\{\cdot\}$  
indicator function

$\min\{\cdot\}$  
minimum function (and $\arg \min$ is for the argument

$\max\{\cdot\}$  
maximum function

$\inf\{\cdot\}$  
inimum function

$Pr(\cdot)$  
probability measure

$\Phi(\cdot)$  
cumulative distribution of a standard normal distribution

$\Phi^{-1}(\cdot)$  
quantile function of the standard normal distribution

$\text{Var}(\cdot)$  
variance

$\{B(t), t \geq 0\}$  
standard Brownian motion stochastic process

$\{X(t), t \geq 0\}$  
stochastic process

$I, J, H, M, M_1, M_2$  
sets of SNP markers

(with small letters denoting the cardinality of the set)

$B$  
Borel sets
Abbreviations

AR attributable risk
AS adaptive sequential design
CA Cochran-Armitage (Trend Test)
CEF conditional error function
CI confidence interval
cM centiMorgan
CPG conditional on parental genotypes
CRP conditional rejection probability
d,D major and minor allele at the DSL
df degrees of freedom
DSL disease susceptibility locus
DNA deoxyribonucleic acid
EPIC European Prospective Investigation into Cancer and Nutrition
FS fixed sample design
FWER familywise error rate
GRR genotype relative risk
GS group sequential design
GWAS genomewide association studies
H₀ null hypothesis
HWE Hardy-Weinberg-equilibrium
LD linkage disequilibrium
m,M major and minor allele at the marker locus
MAF minor allele frequency
MRC Medical Research Council
SNP single nucleotide polymorphism
TDT Transmission/Disequilibrium Test
1 Introduction

“As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality.”


Every copy of the human genome is unique and differs in sequence from any other copy in the population by roughly 1 in 1,250 nucleotides (International HapMap Consortium, 2005; 2003; Sachidanandam et al., 2001). It is this and other variation in DNA (Khaja et al., 2006; Check, 2005) which represents a fossil record of the structure of ancestral populations and which might be the key to understanding phenotypic differences between humans. Facing a worldwide increased prevalence and an anticipated increasing incidence of complex human diseases such as asthma, cardiovascular disease, diabetes or obesity (WHO, 2006; Mathers and Loncar, 2006), identifying links between patterns and determinants of sequence variation and disease risk has become a central goal of human genetics. In order to deal with the uncertainties in human biological reality, mathematical models in statistical genetics have been proposed.

A traditional means of discovering disease genes begins with family-based linkage scans, looking for regions of the genome that tend to be transmitted within families in a manner that parallels the transmission of the trait, followed by a variety of fine-mapping techniques. This approach has been highly successful for mapping major genes responsible for Mendelian disorders in which a single mutation causes the (monogenetic) disease and shows a distinct pattern of cosegregation. Fine-mapping of the putative risk susceptibility loci through linkage analyses, however, will only be feasible with the availability of sufficient recombination events, requiring large pedigrees (Boehnke, 1994).
Since the utility of the linkage approach has also been questioned for elucidating the genetic components of complex, common traits (e.g., Altmüller et al., 2001; Risch and Merikangas, 1996), a considerable effort is being put into the development and conduct of genetic association studies or fine-mapping studies exploiting linkage disequilibrium (LD) between alleles of genetic markers. Recent developments in high-throughput molecular technology (e.g., Syvänen, 2005), the availability of information on genetic variants in databases (e.g., Eberle et al., 2006; International HapMap Consortium, 2005; Hinds et al., 2005), and the initiation of several large-scale, population-based human sample collections (e.g., EPIC (Gonzalez, 2006), MRC/Wellcome Trust Biobank UK (Ollier et al., 2005), Multi-ethnic Cohort (Kolonel et al., 2004)) offers unique opportunities for understanding the (genetic) causes of common diseases.

However, until recently, the results of such investigations in complex traits have been rather disappointing (e.g., Freimer and Sabatti, 2007; Dupuis and O’Donnell, 2007; Todd, 2006). This may be in part due to the fact that these traits are caused by a complex combination of multiple genetic risk factors, environmental exposures, and behaviors. On the other hand, there is evidence that many of the conducted studies did not appropriately address the resulting methodological challenges (e.g., Balding, 2006; Wang et al., 2005; Ott, 2004; Weiss and Terwilliger, 2000). The latter serves as the main motivation for this work.

Consequently, this work will first provide an overview of genetic mapping methods: genetic linkage analysis and linkage disequilibrium mapping/genetic association analysis. As this overview is rather general and primarily serves as a description of the mechanisms exploited by investigators, a subsequent section will focus on some designs frequently used for genetic association studies. Chapter 3 is an introduction to adaptive designs and the methodology used in the later chapters. An often necessary prerequisite to introduce flexible interim design adaptations is the availability of effect size estimators. Thus, both point and interval estimators for genetic effect sizes will be presented or developed (Scherag et al., 2002). Subsequently, it will be shown how these estimators can be implemented in a data-adaptive group sequential procedure for a candidate gene
1 Introduction

Monte Carlo simulations are performed to evaluate the statistical properties of the adaptive procedure in comparison to traditional fixed sample, conventional group sequential designs, and other adaptive designs. The method’s applicability will be demonstrated by a numerical example. Finally, a completely new flexible design for a two-stage genomewide association study is proposed. In this setting the major strength of the adaptive methods will become most obvious - its flexibility to allow for modifications of various aspects of the study like the number of genetic markers evaluated during the course of the study while providing control of the genomewide type I error and using the information already collected.

In addition to cost-efficient multi-stage designs which have been proposed for genomewide association studies (e.g., Müller et al., 2007; Wang et al., 2006), this work shows that methods for data-adaptive, flexible design modifications may further help to improve the allocation of resources to those chromosomal regions which showed the greatest evidence for harboring DNA sequence variation related to disease risk.
2 Genetic mapping of complex traits

This chapter is an introduction to both the genetic terminology and the mechanisms used for gene mapping. A second section deals with the related statistical concepts followed by examples of frequently applied statistical tests for dichotomous traits.

2.1 Biological background

The genetic information of an individual is contained in 23 pairs of chromosomes in the cell nucleus: 22 paired autosomes and two sex chromosomes. Only the former will be considered in this work. Each chromosome consists of a strand of deoxyribonucleic acid (DNA) with units known as base pairs. A chromosomal location i.e., either a single base pair or a collection of consecutive base pairs is termed a genetic locus. At each locus, there may be distinct variants called alleles. For an individual, the pair of alleles at a locus is called the genotype. A genotype is called homozygous if the two alleles are the same allelic variant and heterozygous if they are different allelic variants. If more than one base pair locus is involved, the pattern of alleles for a single chromosome is called a haplotype; together, the two haplotypes for an individual is still called a (multilocus) genotype. Due to their present widespread use for the gene mapping of complex traits, this work focuses on biallelic loci and single nucleotide polymorphisms (SNPs).

Usually to fully describe the genetic model of a trait three components are needed. The first component is called penetrance and denotes the relationship between the (often unobserved) genotypes and the observed individual phenotype which is usually the trait or disease of interest. Thus, the penetrance is the probability of a phenotype given a genotype. A complete genetic model requires specification of the penetrances of all possible genotypes. The transmission probability i.e., the probability with which a particular
allele or haplotype of a parental genotype is transmitted to an offspring, is the second
component of the genetic model. Finally, the third component is the (distribution of)
allele frequencies in the population which is required to determine prior probabilities of
genotypes when inferring genotype from phenotype.

Given a set of phenotypic data on pedigrees, the most likely genetic model can be
estimated using statistical techniques known as segregation analyses (Thomas, 2004).
For most complex traits reduced penetrances can be expected, several genetic loci will
probably influence the trait, and the genetic effect of each locus will be rather small (e.g.,
Botstein and Risch, 2003; Strauch et al., 2003a,b; Weeks and Lathrop, 1995). Furthermore,
the set of loci or alleles influencing the trait may differ between pedigrees which
is referred to as locus or allelic heterogeneity, respectively (Terwilliger and Weiss, 1998).
In such polygenic situations, estimation of the genetic model is difficult to impossible.
Historically, two directions to genetic mapping of complex traits have evolved: those that
require prior specification of a genetic model for the trait being studied study (“model-
based” methods), and those that do not (“model-free” methods). Note that even though
the term “model-free” is often exchangeably used with the term “nonparametric”, nei-
ther of which denote statistical methods that are free of assumptions concerning e.g.,
the class of the underlying probability distributions (Brunner and Munzel, 2002). It has
actually been shown that some “model-free” methods are special cases of “model-based”
methods (e.g., Knapp and Strauch, 2004; Elston, 1998). As it turns out, the difference
between these directions is merely a question of prior-knowledge concerning the range of
alternative hypotheses that will be investigated - if the true genetic model is known, an
appropriate, most powerful “model-based” statistical test can always be chosen (Hodge,
2001). This circumstance will become obvious in section 2.2. For now, the following
two sections deal with the biological mechanisms underlying statistical gene mapping
techniques for complex traits.
2 Genetic mapping of complex traits

2.1.1 Genetic linkage analysis

Cosegregation at two loci denotes the situation where the alleles at two loci that are paternal (maternal) in origin tend to pass to the same gamete (sperm or egg) and, hence, are transmitted together to an offspring. However, during human meiosis, when gametes are formed, the chromosomes pair up together and DNA segments of the paternal and maternal chromosomes interchange by a process known as crossing over. If an odd number of crossovers occurs between two loci, then the alleles at the two loci that an offspring receives from one parent are no longer identical to those that occur in one of the parental chromosomes - they have recombined. Genetic linkage denotes the situation of loci being close together which results in a decreased probability of a recombination event and an increased probability of their cosegregation. The proportion of gametes in which recombination events are expected to occur between two loci is the recombination fraction between them, usually denoted by $\theta$. In other words, $\theta$ is the probability that two loci become recombinant during meiosis. If the two loci are far apart or on different chromosomes, segregation at one locus is independent of that at the other. This is the null hypothesis of linkage analysis with $\theta = 0.5$. As the distance between two loci decreases, $\theta$ tends toward 0. This highly predictable property can be exploited to derive genetic distances in units of Morgan, where one Morgan denotes the distance in which one cross-over per meiosis is expected to occur. A Morgan is divided into centiMorgans (cM), with 1 Morgan = 100 cM.

In linkage analysis (Ott, 1999), the genotypes at so-called genetic marker loci can be measured with a high degree of certainty. Additionally, their physical position in base pairs on the chromosome will be well-known. For a trait or disease susceptibility locus (DSL) the genotypes are unknown. They can only be inferred with varying degrees of accuracy through the trait phenotype. Each single genotyped marker locus or a combination thereof is tested for linkage to the phenotype understudy. In doing so, one seeks to infer the chromosomal regions harboring the (linked) DSL. Since, in genomewide linkage scans, marker loci are spread throughout the genome, at least one of them should be linked to any new gene one wishes to isolate - provided that there is a genetic component.
related to the trait investigated. The closer one can get to the gene, the easier it will be to identify the gene in subsequent labor-intensive molecular analyses. Typically for complex traits, the regions obtained from genomewide linkage scans comprise intervals of 10 cM which corresponds to approximately 10,000,000 base pairs in physical distance (Boehnke, 1994).

2.1.2 Linkage disequilibrium mapping - genetic association analysis

The physical resolution in genetic linkage studies is basically limited by the observed number of recombination and non-recombination events between the marker and the DSL in the sample. Traditionally, fine-mapping of linked regions was followed-up by linkage disequilibrium (LD) mapping / genetic association analyses. This approach has been highly successful in mapping Mendelian traits with a simple mode of inheritance (Botstein and Risch, 2003; Risch, 2000). However, for the detection of DSL with small effects, Risch and Merikangas (1996) argued that LD mapping would be more powerful than “model-free” linkage analysis based on allele-sharing. Ever since (genomewide) association studies were proposed as the future for identifying genetic determinants of complex traits (Jorgenson and Witte, 2006a; Altshuler and Clark, 2005; Wang et al., 2005; Hirschhorn and Daly, 2005; Carlson et al., 2004; Cardon and Bell, 2001). Thus, this section is an introduction to the principles underlying LD mapping.

Again, consider two loci - a marker locus for which the genotypes are given with a high degree of certainty and a DSL which has to be inferred (see Figure 2.1). Originally, there are two alleles at the marker locus: \( M \) and \( m \). This locus is called polymorphic while the DSL is monomorphic i.e., only one allele \( d \) exists at the DSL. At some later time point in population history, a mutation occurs at the DSL and creates a second allele \( D \), which may be considered disease influencing or causing. Since the mutation occurred on the \( M \) background, only three out of four possible gametes will be present in the population: \( DM, dM, \) and \( dm \). At this time allele \( D \) is always found with allele \( M \) on the same chromosome - there is complete LD between the alleles. With passing generations, this allelic association is eroded by recombination events between the loci.
and the genetic association between the alleles is homogenized. Asymptotically linkage equilibrium is reached with the result that the ancestral (disease associated) haplotype is no longer detectable.

The first description of this phenomenon for random-mating populations can be found in Weinberg (1909). Lewontin and Kojima (1960) introduced the term “linkage disequilibrium” and formalized it by a quantitative measure, which is the deviation of observed haplotype frequencies from expected haplotype frequencies. Let $p$ and $q$ denote the allele frequencies of the minor DSL and the minor marker allele. With haplotype probabilities
Table 2.1: Haplotype probabilities for two biallelic loci in relation to their marginal allele frequencies.

<table>
<thead>
<tr>
<th>haplotype</th>
<th>M</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>Pr(DM)</td>
<td>Pr(Dm)</td>
</tr>
<tr>
<td>d</td>
<td>Pr(dM)</td>
<td>Pr(dm)</td>
</tr>
</tbody>
</table>

at the two biallelic loci as depicted in Table 2.1, LD as measured by $\Delta$ is defined as

$$
\Delta = Pr(DM) - pq \\
= -[Pr(Dm) - p(1 - q)] \\
= -[Pr(dM) - (1 - p)q] \\
= Pr(dm) - (1 - p)(1 - q).
$$

Note that $\Delta$ is also a function of the allele frequencies, making it difficult to compare levels of LD. Therefore, several alternative measures with very different properties have been proposed (e.g., reviewed in Pritchard and Przeworski, 2001; Jorde, 2000; Guo, 1997; Devlin and Risch, 1995).

Two important forces governing the decay of LD are the genetic distance between the respective loci and the age of the mutation. Due to the relationship between LD and distance, one may estimate the location of an unknown DSL based on the LD between a disease allele and other known marker alleles. Hence, this is called “indirect mapping” in contrast to “direct mapping” where the marker allele itself is presumably disease related (Zondervan and Cardon, 2004). If a genetic SNP marker is highly correlated with another SNP marker or with a haplotype, it is either called “SNP-tagging” or “haplotype-tagging”, respectively (e.g., Stram, 2004). This idea has stimulated the design of marker arrays that capture most of the common variation within a chromosomal region (e.g., International HapMap Consortium, 2005).

However, in contrast to linkage between loci, the relationship between LD at alleles and physical distance is not that straightforward. Several evolutionary forces like selection, genetic drift or population admixture all have an influence on LD (e.g., Hartl
Table 2.2: Summary of the relationship between linkage and association (adapted from Wienker (priv. comm.)).

<table>
<thead>
<tr>
<th>association</th>
<th>linkage</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>mutations/variants</td>
</tr>
<tr>
<td>yes</td>
<td>- single or few</td>
</tr>
<tr>
<td></td>
<td>- fairly recent</td>
</tr>
<tr>
<td></td>
<td>- closely linked</td>
</tr>
<tr>
<td>no</td>
<td>decay of initial</td>
</tr>
<tr>
<td></td>
<td>LD</td>
</tr>
</tbody>
</table>

and Clark, 2007). As a consequence, assumptions have to be made for gene mapping purposes (e.g., about the age of the founder mutation). The “common disease/common variant” hypothesis (e.g., Peng and Kimmel, 2007; Yang et al., 2005; Pritchard and Cox, 2002; Pritchard, 2001; Terwilliger and Weiss, 1998), upon which most gene mapping efforts in complex traits rely, is one example. Thus, if there is, for instance, some evidence that a certain marker allele is found at a higher frequency in a sample of cases compared to a sample of controls this might indicate the presence of an ancestral disease related haplotype. Admittedly, this allelic association may also be due to other reasons summarized by the expression “spurious association”. An example is confounding due to ethnicity known as “population stratification” in which differences in allele frequency between populations are coupled with differences in other unknown disease risk factors. Special statistical methods have been developed (e.g., Epstein et al., 2007; Köhler and Bickeböller, 2006; Hoggart et al., 2003; Cardon and Palmer, 2003; Satten et al., 2001; Pritchard et al., 2000; Devlin and Roeder, 1999) which deal with this problem. In the majority of cases this affects studies that use unrelated individuals where the transmission probability of the genetic model cannot be accessed. As such “spurious association” findings will usually not be accompanied by evidence of linkage (Lander and Schork, 1994), family-based association studies (discussed below) have been suggested. These studies allow for joint testing of linkage and association.
2 Genetic mapping of complex traits

In a nutshell, association differs from linkage in that the same haplotype is assumed to be associated with the trait in a similar manner across the whole population, while linkage allows for allelic heterogeneity, i.e., different alleles are associated with the trait in different families. A summary of the situation one faces in gene mapping is provided in Table 2.2.

2.2 Statistical methods for genetic association analysis

Designs for genetic association studies are often subdivided into those that use families and into population designs that use unrelated individuals (for a review see Cordell and Clayton, 2005). For dichotomous phenotypes the case-control design is the most widely applied population-based approach. Its advantages are that large samples can be obtained relatively easy and that it is straightforward to compare the marker genotype or allele frequencies in the cases to those observed in the controls. In the absence of systematic errors/biases, significant differences in distributions might indicate a genetic association either directly or indirectly due to LD. The results of a case-control study for one SNP can be summarized in a 2 x 3 frequency table for genotypes or in a 2 x 2 frequency table for alleles. Sasieni (1997) made some recommendations to analyze these tables. Especially for the case of violations of the Hardy-Weinberg-equilibrium (HWE), when each individual’s two chromosomes can no longer be viewed as independent, only genotype-based tests lead to valid results (Clayton, 2001; Schaid, 1998). Restricting the examination of the case-control approach to genotype-based analyses, the marker genotype probabilities in cases can be summarized as depicted in Table 2.3 using the established notation. In addition, a penetrance row vector $f = (f_0, f_1, f_2)$ for carriers of zero, one, or two copies of the disease-related marker allele $M$ is introduced with $f_i = Pr(case|i M$ alleles in the genotype) with $i = 0, 1, 2$.

Hence, the marker genotype frequencies in cases can be described by a multinomial distribution with parameters $(\varphi_0^{case}, \varphi_1^{case}, \varphi_2^{case})$, where the subscript corresponds to the number of high risk alleles in the genotype. For controls a table similar to Table 2.3 can be constructed with $f_i$ replaced by $(1 - f_i)$ for $i = 0, 1, 2$ and the respective multinomial
Table 2.3: Marker genotype probabilities for cases in relation to marker-disease haplotype probabilities shown in Table 2.1.

<table>
<thead>
<tr>
<th>marker genotype</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>mm</td>
<td>( \text{Pr}(\text{dm})^2 f_0 + 2\text{Pr}(\text{dm})\text{Pr}(\text{dM}) f_1 + \text{Pr}(\text{dM})^2 f_2 )</td>
</tr>
<tr>
<td>mM</td>
<td>(2\text{Pr}(\text{dm})\text{Pr}(\text{Dm}) f_0 + 2[\text{Pr}(\text{dm})\text{Pr}(\text{DM}) + \text{Pr}(\text{dM})\text{Pr}(\text{Dm})] f_1 + 2\text{Pr}(\text{dM})\text{Pr}(\text{DM}) f_2 )</td>
</tr>
<tr>
<td>MM</td>
<td>( \text{Pr}(\text{Dm})^2 f_0 + 2\text{Pr}(\text{Dm})\text{Pr}(\text{DM}) f_1 + \text{Pr}(\text{DM})^2 f_2 )</td>
</tr>
</tbody>
</table>

distribution parameters \((\bar{\varphi}_0, \bar{\varphi}_1, \bar{\varphi}_2)\). Consequently, the null hypothesis of no association can be written as \(\varphi_i^{\text{case}} = \varphi_i^{\text{control}}\) for \(i = 0, 1, 2\). Note that the related statistical test will be a 2 degrees of freedom (df) test though more powerful 1 df tests may be employed (e.g., Chiano and Clayton, 1998). Sasieni (1997) proposed to use the “Cochran-Armitage (CA) Trend Test” (Armitage, 1955; Cochran, 1954) for investigating arbitrary trends related to genotype frequency differences in cases and controls. As an example of statistical tests used for the analysis of case-control data it will be briefly presented here (for more details see Zheng and Gastwirth, 2006). Assume that \(w_i\) for \(i = 0, 1, 2\) is the (linear) trend of interest under the alternative hypothesis. According to Zheng et al. (2003) the choice of \(w = (0, 0.5, 1)\) will yield an asymptotically local optimal test if the true genetic model is additive i.e., a linear allele-dose response relationship is assumed. Notice the smooth transition from a “model-free” to a “model-based” test.

Provided that \(n\) denotes the total sample size, let \(n^{\text{control}}\) denote the sample size of the cases and let \(N_i^{\text{control}}\) for \(i = 0, 1, 2\) be the random variables for the genotype counts. For the cases a similar notation holds respectively and \(\phi = n^{\text{case}}/n\) is their proportion relative to the total sample. Ignoring its dependence on \(w_i\) in the notation, the test statistic of the CA trend test can be written as

\[
T_{CA} = \frac{U}{\sqrt{\text{Var}(U)}} \quad \text{with} \quad U = \sum_{i=0}^{2} w_i [(1 - \phi)N_i^{\text{case}} - \phi N_i^{\text{control}}].
\]

A consistent estimate of \(\text{Var}(U)\) under both the null and the alternative hypothesis is
2 Genetic mapping of complex traits

given by

\[
\widehat{\text{Var}}(U) = n\phi(1-\phi)^2 \left\{ \sum_{i=0}^{2} w_i^2 \frac{N_{\text{case}}^i}{n_{\text{case}}} - \left[ \sum_{i=0}^{2} w_i^2 \frac{N_{\text{case}}^i}{n_{\text{case}}} \right]^2 \right\} \\
+ n\phi^2(1-\phi) \left\{ \sum_{i=0}^{2} w_i^2 \frac{N_{\text{control}}^i}{n_{\text{control}}} - \left[ \sum_{i=0}^{2} w_i^2 \frac{N_{\text{control}}^i}{n_{\text{control}}} \right]^2 \right\}.
\]

Based on the central limit theorem, the asymptotic distribution of \( T_{CA} \) is \( \mathcal{N}(0,1) \) under the null hypothesis of no association.

The most widespread test for family-based association studies is the "Transmission/Disequilibrium Test" (TDT) which is applicable to genotype data from so-called trios - an affected offspring and her or his two biological parents (Spielman and Ewens, 1996; Spielman et al., 1994, 1993; Ott, 1989; Falk and Rubinstein, 1987). Unlike in the former case of the CA trend test, however, the TDT is an allele-based test. It compares the observed number of alleles that are transmitted from heterozygous parents to their affected child with those expected under Mendelian inheritance (see Table 2.4). The assumption of Mendelian transmissions is all that is needed to ensure valid results of the TDT and related approaches. An excess of alleles of one type among the affected offspring indicates that a DSL for the phenotype of interest is linked with the marker locus and that alleles at the DSL and the marker locus are associated. Originally, the TDT was used to test for linkage in the presence of an association. However, because both linkage (\( \theta < 0.5 \)) and linkage disequilibrium (\( \Delta > 0 \)) have to be present for the TDT to reject the null hypothesis (\( \Delta(\theta - 0.5) = 0 \)), the TDT is now typically used as a test for association (Ewens and Spielman, 2005). The dual-alternative hypothesis also implies that the TDT avoids situations of "spurious associations". This property has stimulated multiple refinements (e.g., evaluated/reviewed in Nicodemus et al., 2007; Laird and Lange, 2006).

Restricting the consideration to the marker allele \( M \), let \( N^T \) be a random number of transmitted \( M \) alleles among \( 2N \) chromosomes of the \( N \) heterozygous parents and let \( N^{NT} \) be the number of non-transmitted \( M \) alleles, respectively. Note that in Table 2.4 these values are related to the two cells that bear on \( \theta \). Furthermore, if \( \theta = 0.5 \), the expected values \( E(N^T) = E(N^{NT}) \), whatever the values of \( p, q, \) and \( \Delta \). When \( \theta = 0.5 \),
Table 2.4: Probabilities of combinations of transmitted and non-transmitted marker alleles $M$ and $m$ among $2n$ parents of affected offspring (adapted from Ott, 1989).

<table>
<thead>
<tr>
<th>Transmitted Allele</th>
<th>Non-Transmitted Allele</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>$q^2 + q\Delta/p$</td>
<td>$q$</td>
</tr>
<tr>
<td></td>
<td>$q(1-q)$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$(1-\theta-q)\Delta/p$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$(1-\theta)\Delta/p$</td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td>$q(1-q) + (\theta-q)\Delta/p$</td>
<td>$1-q$</td>
</tr>
<tr>
<td></td>
<td>$(1-q)^2 - (1-q)\Delta/p$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$-(1-\theta)\Delta/p$</td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>$q + \theta\Delta/p$</td>
<td>$1-q-\theta\Delta/p$</td>
</tr>
<tr>
<td></td>
<td>$1-q - \theta\Delta/p$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$1$</td>
<td></td>
</tr>
</tbody>
</table>

The contributions from two heterozygous parents are independent and a test statistic can be derived as

$$T_{TDT} = \frac{U}{\sqrt{\text{Var}(U)}} \quad \text{with} \quad U = N^T - N^{NT} \quad \text{and} \quad \text{Var}(U) = N^T + N^{NT}.$$  

Under the null hypothesis of no linkage and no association, $T_{TDT} \sim \mathcal{N}(0,1)$ asymptotically. Note that this test statistic will be called $T_{TDT}$ even though it is not the classical TDT/McNemar test statistic. The test statistic of the classical TDT is given by $(T_{TDT})^2$ as the TDT is usually performed as a two-sided test. $T_{TDT}$, however, is more readily applicable to the problem of one-sided alternatives i.e., the situation of using prior knowledge of $M$ referring to the disease related haplotype. Given such a situation, assume that $\alpha$ was set to 0.025 (one-sided) and that an appropriate number of trio genotype data had been collected. The null hypothesis of no genetic linkage and no allelic association can be rejected if $T_{TDT} \geq z_{1-\alpha}$ provided that the right term denotes the $(1-\alpha)$-quantile of a standard normal distribution and that the direction of the effect is consistent with the pre-specified direction. In summary, this chapter first provided an introduction to the biological terminology and the principles underlying gene mapping. Second, it was demonstrated how these mechanisms can be used to develop test statistics for dichotomous phenotypes in genetic association studies which are frequently used in practice. Both statistics will also be used in the subsequent chapters.
3 Mathematical background

Gene mapping studies are often conducted in a rather explorative way with conclusions drawn that rarely reflect this hypothesis-generating process. Most obviously in genetic association studies, this has resulted in multiple failures to replicate initial reports (e.g., Ott, 2004; Wacholder et al., 2002; Hirschhorn et al., 2002) - presumably instances of false positive findings (e.g., Moonesinghe et al., 2007; Ioannidis, 2005; Wacholder et al., 2004; Weiss and Terwilliger, 2000). As a consequence, guidelines (for publications) have either been propagated or proposed (e.g., Ioannidis, 2007; Hattersley and McCarthy, 2005; Freimer and Sabatti, 2005; Little et al., 2002; Cooper et al., 2002), and they can be expected to become increasingly important in future research. Control of the type I error rate and adequate sample size considerations are major topics of these proposals. Both sections of this chapter are related to these fundamental requirements.

First, a general framework of hypotheses testing is provided; followed by a review of adaptive designs that allow for controlling the type I error. In addition, by modifying the initial planning, these ideas also comply with the researchers interest in the study’s power to detect disease related sequence variations which will be demonstrated in chapter 4 and 5. A separate subsection is dedicated to a general principle for design adaptations (Müller and Schäfer, 2004) which is of major importance for the subsequent chapters. The second section of this chapter deals with point and interval estimates of genetic effects. These estimators will be used to justify sample size modifications for candidate gene association studies as outlined in chapter 4.
3 Mathematical background

3.1 Adaptive designs and adaptive design methods

3.1.1 General framework

In general, statistical inference deals with using a set of observations - values of random variables $X$ - to obtain information concerning the distribution of the $X$. Sometimes assumptions about the distribution class of the $X$ can be made - certain events may be uniformly distributed or may be stochastically independent. In the case of testing single point hypotheses, it is for instance implied that the true distribution of the $X$ can be quantified by a parameter $\theta$. Testing statistical hypotheses and type I error rate control at a level $\alpha$ requires the formulation of a null hypothesis (e.g., $\theta \leq \theta_0$) and an alternative hypothesis (e.g., $\theta > \theta_0$ for a one-sided alternative). Next, a level $\alpha$ to control the type I error rate and an appropriate test statistic $T$ will be chosen. After sampling a pre-specified number of observations a decision about whether to reject or not to reject the null hypothesis will be made based on the value of the test statistic $T$ derived from the data. While this roughly summarizes the requirements for control of the type I error rate and hypothesis testing in a standard situation for a single data set, assume that the same null hypothesis has been investigated in $k(\geq 2)$ independent data sets. Suppose that one wants to combine the data sets in order to test the overall null hypothesis which is the intersection null hypothesis $H_0 = \bigcap_{i=1}^{k} H_{0i}$ implying $\theta_i \leq \theta_0$ for each $i = 1, \ldots, k$ against the alternative hypothesis $H_1$ that at least one $\theta_i > \theta_0$. This can be achieved by combining independent test statistics $T_i$ related to the parameter of interest from the $i^{th}$ data set by means of a combination function. If data sets are in fact independent studies, this represents a somewhat simplified meta-analysis scenario (e.g., Hedges and Olkin, 1985).

The basic idea of so-called adaptive designs was to understand a new, prospectively conducted study as a *pre-planned* meta-analysis of two or more substudies which are usually called stages (Bauer and Köhne, 1994; Bauer, 1989). For adaptive designs, however, the type I error rate can and has to be controlled. This is achieved if the choice of an appropriate combination function (including the number of stages), the
distribution of every $T_i$ under $H_0$, and the design and sample size of the first stage are specified before the start of the study. Furthermore, in case of a single point null hypothesis, both stochastic independence of the $T_i$ and maintenance of the initially planned marginal distributions of every $T_i$ under $H_0$ irrespective of design changes are assumed. If these requirements are fulfilled, investigators are allowed to modify the number of subjects entering the $i^{th}$ stage (for $i > 1$) as well as the statistical test used in this stage after completion of the $(i-1)^{th}$ stage based on arbitrary information. Moreover there is an additional requirement, which is common to all the proposals presented below: at least the option for changes has to be specified before the study starts in order to ensure control of the type I error rate. This is why they will be called “adaptive design”.

A greater challenge is the preservation of the type I error rate if the eventuality of making design changes was not foreseen in the original plan. A statistical principle that enables investigators to conduct such “design adaptations” is presented in section 3.1.3.

### 3.1.2 Adaptive designs

The introduction of adaptive designs for clinical trails (Bauer and Köhne, 1994; Bauer, 1989) is viewed as a milestone in the general development of statistical methods for sequential statistical analyses (e.g., Schäfer et al., 2006; Koch, 2006). Before this, no data-dependant design changes were allowed during the course of the trial even though some flexibility was, for instance, provided by the alpha spending approach (e.g., DeMets and Lan, 1994). Ever since both methodological developments and their applications have rapidly grown (Bauer and Einfalt, 2006). Therefore, this section cannot be an exhaustive presentation. Particularly, an illustration of point and interval estimates (e.g., Brannath et al., 2006) as well as a discussion of ideas summarized under the topic “internal pilot study” (see Friede and Kieser, 2006) will be omitted.

Using Fisher’s combination rule (Fisher, 1932), Bauer (1989) and Bauer and Köhne (1994) developed the first adaptive designs. Given that $p_i$ denotes the observed one-sided p-value from a continuous test statistic used to test the individual $H_{0i}$ in the $i^{th}$-stage with $1 \leq i \leq k$, an $\alpha$ level test rejects the intersection null hypothesis $H_0 = \bigcap_{i=1}^{k} H_{0i}$ if
Figure 3.1: Rejection region (grey; \( \alpha = 0.2 \)) on probability scale for Fisher’s inverse \( \chi^2 \) test with two stages.

\[
\prod_{i=1}^{k} p_i \leq \exp \left( -\frac{1}{2} \chi_{2k,1-\alpha}^2 \right)
\]

where \( \chi_{2k,1-\alpha}^2 \) denotes the \((1-\alpha)\)-quantile of a chi-squared distribution with \( 2k \) df. This results because under the intersect null hypotheses the \( p_i \) are known to be independent uniformly \( U(0,1) \) distributed and minus the logarithmized p-values are \( \frac{1}{2} \cdot \chi_{2}^2 \)-distributed. Note that Fisher’s method also keeps the global level when the p-values under \( H_0 \) are stochastically larger than the uniform distribution (Wassmer, 2001, chap. 3). Figure 3.1 is a graphical representation for a design with two stages and p-values \( p_1 \) and \( p_2 \). The range of \((p_1, p_2)\) is the unit square \([0,1] \times [0,1]\) and an \( \alpha \) level test rejects \( H_0 \) if \((p_1, p_2)\) is located within the critical region (see Figure 3.1). Note that the test will always reject \( H_0 \) if either \( p_1 \leq c_\alpha \) or \( p_2 \leq c_\alpha \) with \( c_\alpha = \exp(-\frac{1}{2} \cdot \chi_{4,1-\alpha}^2) \) denoting the critical value. Thus, if \( p_1 \leq c_\alpha \), there is no need to conduct the second stage of the study. This property is called “non-stochastic curtailment”. Now suppose that after observing \( p_1 \) of the first stage, it is decided to run two instead of one additional stages with corresponding p-values \( p_{2a} \) and \( p_{2b} \). Then \( p_2 \) is defined as a solution to \(-2 \ln(p_{2a}p_{2b}) = \chi_{4,1-p_2}^2\), where the left term denotes the \((1-p_2)\)-quantile of a chi-squared distribution with 4 df. Finally, applying
the original rule, $p_1$ and $p_2$ will be combined. Obviously, for $k = 3$ stages planned in advance, the critical value for rejecting $H_0$ would have been different. This underlines the necessity of pre-specifying the number of stages $k$ and the combination function to obtain a valid level $\alpha$ test. Moreover, it is an example of a recursive application of Fisher’s combination test. A generalization of this idea was proposed by Brannath et al. (2002).

Admittedly, other divisions of the unit square in Figure 3.1 that meet the $\alpha$ level criterion are conceivable. For instance, Bauer and Köhne (1994), and Wassmer (1999) for more than two stages, modified Fisher’s combinations rule by explicitly including the option for early stopping with acceptance of $H_0$ - the so called “futility stops”. Let $\alpha_0$ and $\alpha_1$ denote the first stage decision boundaries for rejection of $H_0$ if $p_1 \leq \alpha_1$ ($\alpha_1 > c_\alpha$) or for acceptance of $H_0$ if $p_1 \geq \alpha_0$. In case of $\alpha_1 < p_1 < \alpha_0$ the second stage will be conducted and $H_0$ can be rejected if $p_1 \cdot p_2 < c_\alpha$. The overall probability for a type I error is given by

$$\int_{\alpha_1}^{\alpha_0} dp_1 + \int_{0}^{c_\alpha/p_1} dp_2 dp_1 = \alpha_1 + c_\alpha (\ln \alpha_0 - \ln \alpha_1),$$

and a level $\alpha$ test can be constructed by finding solutions for the case that this term equates $\alpha$.

While the work on chi-square-based combination functions was further extended (Hartung and Knapp, 2003), Proschan and Hunsberger (1995) provided an alternative view of combination functions for adaptive two-stage designs in the Gaussian distribution model with known variance by introducing the concept of conditional error functions (CEF). Any non-increasing function $\alpha(p)$ with range $[0, 1]$ that satisfies

$$\int_{0}^{1} \alpha(p_1)dp_1 = \alpha$$

is called a conditional (type I) error function. The intersection null hypothesis $H_0$ is rejected if and only if the second stage $p$-value, $p_2$ is smaller than $\alpha(p_1)$. In other words, the hypothesis test for the second stage is carried out at a level $\alpha(p_1)$. Thus,
the design of the second stage does not have to be specified prior to the results of the first stage and as long as the significance level $\alpha(p_1)$ is used, the type I error rate will be controlled. Even though Proschan and Hunsberger (1995) propagated the use of their “circular” conditional error function, a comparison by Wassmer (1998) revealed that there is virtually no difference between the Bauer and Köhne version of Fisher’s combination test (Bauer and Köhne, 1994) and the “circular” conditional error function (Proschan and Hunsberger, 1995) in terms of statistical power and average sample size.

One class of combination functions/conditional error functions, however, deserves special attention due to notably statistical properties - combination functions that are based on sums of (weighted) inverse normal deviates. Define $Z_i = \Phi^{-1}(1 - p_i)$ for $i = 1, \ldots, k$, where $\Phi^{-1}(\cdot)$ denotes the quantile function of the cumulative distribution function for a standard normal distribution. Under $H_0$, each $Z_i \sim \mathcal{N}(0, 1)$ and an $\alpha$ level test rejects $H_0$ if

$$\frac{1}{\sqrt{k}} \sum_{i=1}^{k} Z_i \geq z_{1-\alpha},$$

where $z_{1-\alpha}$ denotes the $(1 - \alpha)$-quantile of a standard normal distribution. A general-
ization is the weighted inverse normal method whereby $H_0$ is rejected at level $\alpha$ if
\[ \sum_{i=1}^{k} w_i Z_i \geq z_{1-\alpha}, \] (3.1)
with fixed weights $w_i$ chosen to satisfy $\sum_{i=1}^{k} w_i^2 = 1$. In this case, the left side of equation (3.1) will be $\mathcal{N}(0,1)$-distributed under $H_0$. Figure 3.2 is an example of the “inverse normal method” for $k = 2$ stages and weights $w_1 = w_2 = \sqrt{0.5}$. Unlike in Figure 3.1, no early stopping of the whole study with rejection of $H_0$ by “non-stochastic curtailment” will be possible. However, Fisher (1998) and Shen and Fisher (1999) implicitly address this issue noting that both the number of stages as well as the weights in an “inverse normal”-representation of the problem do not have to be selected in advance. Instead both may be random variables which can be specified based on data collected in earlier stages of the study leading to what Fisher calls a “self-designing” trial. To ensure control of the type I error rate the basic requirement is that the standardized total variance of the final test of 1 is distributed over the different stages i.e., it is required that $\sum_{i=1}^{k} w_i^2 = 1$ almost surely. The “variance spending”-method starts with a weight for the first stage e.g., calculated from the initial design parameters presumably weighting all observations equally (similar to Figure 3.2). The weight function at the $j^{th}$ stage with $j > 1$ is estimated based on the accumulated data prior to the $j^{th}$ stage and as long as $\sum_{i=1}^{j} w_i^2 < 1$ the procedure will continue. When the weight function is used up at the $k^{th}$ stage, i.e., $\sum_{i=1}^{k} w_i^2 = 1$, the study will be stopped and $H_0$ will be rejected if $\sum_{i=1}^{k} w_i Z_i \geq z_{1-\alpha}$ with the weights derived from the procedure. Regardless of its practical appeal like not requiring numerical integration, the method has some limitations. A minor remark is that the option for early stopping with acceptance of $H_0$ was not used for adjusting their final rejection region which will result in a loss of power. A major drawback becomes obvious when imagining the situation that a large effect has been observed during the course of a study, and so far only a small amount of the total weight has been spent. At least one more stage has to be sampled before a rejection of $H_0$ is possible. In this case the data dependent choice of weights may lead to a rather dangerous strategy - the remaining large weight might be given to the last stage which will presumably comprise
3 Mathematical background

a small sample size (for critical reviews see Burman and Sonesson, 2006; Jennison and Turnbull, 2003; Tsiatis and Mehta, 2003).

Alternatively, Cui et al. (1999) and Lehmacher and Wassmer (1999) suggested to use the classical group sequential methodology (reviewed in Wassmer, 2001; Jennison and Turnbull, 1999) where both early stopping with rejection and acceptance of null hypotheses had been extensively dealt with. The application of efficient group sequential tests (Tsiatis and Mehta, 2003) and boundaries derived from e.g., the recursive convolution algorithm (Armitage et al., 1969) becomes obvious if the partial sums \( \sum_{i=1}^{j} w_i Z_i \) for \( j < k \) are described as finite dimensional marginal distributions of a Brownian motion process, monitored at so-called “discrete information time points”. This, however, requires some introduction to Brownian motion stochastic processes (for more detail see Lan and Zucker, 1993; Whitehead, 1999). A stochastic process \( \{X(t), t \geq 0\} \) is said to be a Brownian motion process, or simply Brownian motion, with a drift parameter \( \delta \) and a variance parameter \( \sigma^2 \) if:

- \( X(0) = 0 \),

- \( \{X(t)\} \) has stationary and independent increments, and

- for every \( t > 0 \), \( X(t) \sim \mathcal{N}(\delta \cdot t, \sigma^2 \cdot t) \).

The process defined by

\[
B(t) = \frac{X(t) - \delta \cdot t}{\sigma}
\]

is a standard Brownian motion with drift 0 and variance parameter 1. Conversely, if \( \{B(t), t \geq 0\} \) is a standard Brownian motion then by defining \( X(t) = \delta \cdot t + \sigma \cdot B(t) \) the stochastic process \( \{X(t), t \geq 0\} \) is a Brownian motion with a drift parameter \( \delta \) and a variance parameter \( \sigma^2 \).

Applied to sequentially collected data, there obviously is no continuous time \( t \). Nevertheless, a discrete “information fraction” or “information time” may be constructed. The proportion \( t \in [0,1] \) of an already assessed sample relative to the total sample size \( n \) is such a measure. A more general information-based framework for study designs
3 Mathematical background

can, for example, be found in Mehta and Patel (2006). For the purpose of this work, it will suffice to show how normally distributed test statistics can be linked to the Brownian motion representation (as proposed by Müller and Schäfer, 1999). Suppose that $T_{t.n} \sim N(\sqrt{t \cdot \bar{n}} \cdot \mu, \sigma^2)$ denotes a test statistic calculated for a sample size of $t \cdot n$. Assume that under $H_0$: $\mu = 0$ and $\sigma = 1$ resulting in a standardized test statistic. Provided that the conditions for a Brownian motion hold, a Brownian motion representation of this test statistic is given by

$$T(t) = \sqrt{t} \cdot T_{t,n} \text{ with } t \in [0, 1],$$

where $T(t)$ is a Brownian motion with a drift parameter $\delta = \sqrt{t} \cdot \mu$ and a variance parameter $\sigma^2$. This representation will be used in chapter 4. Note that power and sample size calculations require a choice of $\delta$ and $\sigma^2$ under the alternative hypotheses.

Obvious from this Brownian motion excursion and fundamental within the classical group sequential design framework, sample sizes of the stages are pre-planned and are not allowed to depend on previous stages - otherwise no control of the type I error rate will be possible (Schäfer et al., 2006; Proschan et al., 1992). Nonetheless, Cui et al. (1999) and Lehmacher and Wassmer (1999) proposed to start a study with an efficient group sequential plan. In case of design-adaptations both the rejection boundaries and the configuration of the pre-planned analyses remain unchanged. However, the observations sampled after the design adaptation are given a different weight than those before the adaptation. Technically both proposals describe the same situation but they are differently motivated. While Cui et al. (1999) face the practical situation of having started with a group sequential plan and requiring a design modification, Lehmacher and Wassmer (1999) use the “inverse normal” combination function to construct an adaptive design. Hence, their design has properties of a group sequential design.

An approach more similar to the idea of Cui et al. (1999) was proposed by Müller and Schäfer (2001) who show that the efficiency advantages of group sequential designs can be easily combined with adaptive designs based on a general framework. The underlying principle will be described in the next section.
3.1.3 Methods for design adaptation

As outlined before, adaptive designs make it possible to start the trial with the design of the first stage only. They require a specification of the sample size for the first stage and a conditional error function to be applied after completion of the first stage (Bauer and Brannath, 2004; Brannath et al., 2002). In a quite different setting one may want to start with a completely specified optimal design such as a conventional group sequential test or a fixed sample design and a method for design adaptations may be applied if necessary. In the case that no design alterations are required, there will be no loss of efficiency as compared to the optimal test.

Müller and Schäfer (2004) have introduced the conditional rejection probability (CRP) principle as a general theoretical instrument for design modifications at any time during the course of a study. Using this principle, all types of design modifications known from adaptive designs can also be implemented in every standard design without inflation of the type I error rate. Moreover, no price has to be paid for the mere option of design modifications. In contrast to adaptive designs, where the specification of a conditional error function is a fundamental design element, the CRP principle uses the “natural” conditional error function which is implicitly defined by the choice of an initial design. The situations considered in the subsequent chapters all assume normally distributed outcomes with known variance. This choice is motivated by a wide scope of applications if large sample approximations and the Brownian motion model can be adopted (Müller and Schäfer, 2001; Schäfer and Müller, 2001; Müller and Schäfer, 1999). The CRP principle, however, is more general as recently demonstrated in the situation of unknown variances and t-distributed test statistics (Timmesfeld et al., 2007).

The CRP principle can best be understood in the context of decision functions. Let $X = (X_1, \ldots, X_n)$ denote the random vector modeling the data collected during the complete study. Again, the objective of the study is to draw inference on the unknown parameter $\vartheta$ within a hypothesis testing framework. Assume that a null hypothesis $H_0 \subset \Theta$ is to be tested using a statistical test formalized by a decision function $\varphi(X)$ with a real interval range $[0,1]$. At the end of the study, if $\varphi(X) = 1$ is realized, $H_0$ will
be rejected. Otherwise, for $\varphi(X) < 1$ and in case of a non-randomized decision function, $H_0$ will be accepted. For a fixed sample design and symmetrical rejection limits $\pm c$, a real valued test statistic $T$ may be calculated for $X$ and the decision function is given by the indicator function $I\{T < -c \text{ or } T > c\}$.

The random vector $X$ may be decomposed in two components $X = (X_1, X_2)$. $X_1$ denotes the component of $X$ up to an interim data inspection and $x_1$ is its realization. $X_2$ is the random vector modeling the future course of the study according to the initial plan. The decision function may be written as $\varphi(X_1; X_2)$. A concrete version of the CRP is the conditional expectation of the decision function given the observed data of the vector $X_1$

$$CRP^\vartheta(x_1) = E^\vartheta(\varphi(X_1; X_2)|X_1 = x_1).$$

After a design adaptation, the data of the new design can be modeled as a random vector $Y$, i.e., $X_2$ will be replaced by $Y$. Furthermore, a new decision function $\psi_{x_1}(\cdot)$ may be defined. In order to control the type I error rate for the design as a whole, the sufficient condition is that the new decision variable holds the type I error risk $CRP^\vartheta(x_1)$ for every $\vartheta \subset H_0$, i.e., the new decision function must fulfill the fundamental inequality

$$E^\vartheta(\psi_{x_1}(Y)|X_1 = x_1) \leq CRP^\vartheta(x_1) \text{ for every } \vartheta \subset H_0.$$

For many practical applications, it will be sufficient to consider special values of $\vartheta$, as for example $\vartheta_0$ in the introduction of this chapter.

The CRP principle will be used in chapter 4 to modify the sample size in a candidate gene study either for a family-based (simulation) or a case-control design (application). In chapter 5 it will be extended to allow for testing of multiple hypotheses in genomewide association studies. Again, both case-control (simulation) as well as family-based (application) scenarios will be outlined. More details about the CRP principle, including a proof that the procedure holds the pre-specified type I error level, can be found in Müller and Schäfer (2004).
3.2 Point and interval estimates of genetic effects

For reliable sample size calculations which yield a desired power, all of the proposed designs for association studies of complex diseases critically depend on a precise prespecification of the genetic effect. In general, little is known about the genetic effect sizes in initial studies (Lohmueller et al., 2003; Ioannidis et al., 2003; Hirschhorn et al., 2002; Ioannidis et al., 2001). Thus, the general idea is to use estimators of genetic effect sizes and allele frequencies derived at an interim analysis in order to justify sample size choices done with a procedure allowing for design modifications. As a consequence, estimators of genetic effect sizes and allele frequencies are required to support the decision to modify the initial design. In chapter 4 these estimates will be used for sample size modifications in a candidate gene association.

While for the case-control design such estimates are similar to those derived from classical epidemiology, the situation for the family-base trio design is more complicated. Consequently, the former will only be briefly described and the primary focus will be on the latter. Moreover, new easy to use formulae to calculate confidence intervals for genotype relative risks (GRRs) of homozygous and heterozygous allele carriers for the population frequency of the risk alleles and for the attributable risk have been contributed by the author of this thesis (Scherag et al., 2002). These formulae are based on likelihood methods.

3.2.1 Case-control design

In classical epidemiology various study designs have been developed to obtain accurate and unbiased estimates of disease risk related to some risk factor. These ideas have been adopted to genetic association studies (Balding, 2006).

For a case-control genetic association study, the objective is to compare the presence of a certain genotype between affected individuals and unaffected controls, who have been selected from the same population as the cases, in order to find associations between the genotypes and the disease. Appropriately conducted i.e., assuming that biases have been avoided, the standard measure of effect in these studies is the odds ratio, defined

36
as the odds of exposure among cases divided by the odds of exposure among controls (for an introduction see Rothman and Greenland, 2005; Benichou and Palta, 2005).

Originally, the odds ratio was proposed as an approximation of relative risk if the disease being studied is relatively rare (prevalence < 10%) given the situation that incident or prevalent cases and controls were sampled cross-sectionally (Greenland and Rothman, 1998). However, it has been shown that depending on the ascertainment strategy and the aim of the study, the odds ratio might as well serve as an approximation to the relative risk for common diseases e.g., if sampling for the case-control study mimics that of a prospective cohort study (Rodrigues and Kirkwood, 1990). In addition, more complex modeling and estimation in genetic case-control studies can be done within a generalized linear regression framework (for a review see Clayton, 2001).

### 3.2.2 Family-based trio design

Unlike for independent individuals, unbiased estimation of genetic effects within family-based designs requires taking into account dependencies between family members. Schaid and Sommer (1993) have developed the likelihood theory for the case parent trio design and have introduced two basic models, a likelihood conditional on parental genotypes (abbreviated CPG) and an unconditional likelihood which is valid under the assumption of Hardy-Weinberg equilibrium (HWE likelihood). The parameters of these models are the allele frequencies and the GRRs for homozygous and heterozygous bearers of the risk allele. Knapp et al. (1995) derive explicit formulae for the maximum likelihood estimators of these parameters. Schaid (1999) investigates the CPG likelihood under additive, dominant, recessive and multiplicative genetic models. Under these genetic models, the GRRs for homozygous and heterozygous bearers of the risk allele are functions of only one risk parameter for which Schaid (1999) derives exact confidence intervals. He also gives an implicit equation for exact simultaneous confidence intervals for both GRRs without assuming any otherwise specified mathematical relationship between the two. However, the solution of this equation is computationally intensive and no computer code is available. Below asymptotic confidence intervals for the GRRs, for the allele fre-
quencies and for the attributable risk in the unconditional HWE likelihood model will be presented. Their properties are investigated by a simulation study. The confidence intervals can be calculated directly from the familial genotype frequencies without the need for numerical methods. No constraints are imposed on the GRRs. In addition, a method for the calculation of CPG confidence intervals will be described and a C++ computer program is provided in the Appendix. Finally, exact Clopper-Pearson confidence intervals (Clopper and Pearson, 1934) for the allele frequency and the heterozygous GRR are presented. They may be especially useful in situations of small (minor) allele frequencies or small sample sizes.

**Point and interval estimates for genetic parameters**

For simplicity only two alleles \(D\) and \(d\) at the candidate gene DSL locus will be considered, and again \(D\) is the risk related allele. The parameters of interest are the allele frequencies \(p\) and \((1 - p)\) of the alleles \(D\) and \(d\), respectively, and the GRRs \(\Psi_1\) and \(\Psi_2\) for heterozygous and homozygous bearers of the risk allele \(D\), respectively, which are defined by \(\Psi_1 = \frac{\text{Pr(affected|Dd)}}{\text{Pr(affected|dd)}}\) and \(\Psi_2 = \frac{\text{Pr(affected|DD)}}{\text{Pr(affected|dd)}}\). From these parameters, the attributable risk \(AR\) (Schaid and Sommer, 1993) can be derived which is defined as the percentage of the disease attributable to the risk allele and is given by

\[
AR = \frac{(\Psi_2 - 1)p^2 + (\Psi_1 - 1)2p(1 - p)}{\Psi_2p^2 + \Psi_12p(1 - p) + (1 - p)^2}.
\]

The data from a case parent trio study can be summarized as frequencies of observed familial genotypes as indicated in Table 3.1 (Schaid and Sommer, 1993). When families are sampled for affected offspring, the conditional probability of sampling a family of the respective genotype is given by the term in the last column of the table. Here the assumption of HWE is made. For ease of notation, the following numbers derived from the observed frequencies in the table are defined. The total number of families in the sample is denoted by \(n\). \(N_{ij}\) is the number of families with the genotype defined by the pair of indices \(i\) and \(j\). Here, \(i\) is the number of \(D\) alleles in the parental generation

38
(3 Mathematical background

(i = 0, \ldots, 4) and \( j \) is the number of D alleles transmitted to the offspring (\( j = 0, 1, 2 \)). \( N_0, N_1, N_2 \) are the numbers of families with 0, 1 and 2 copies of the D allele transmitted to the offspring and \( K \) is the number of copies of the D allele among the 2N alleles not transmitted to the offspring. Note that these numbers may be derived from the observed frequencies defined in Table 3.1 by

\[
N_0 = N_{00} + N_{10} + N_{20}, \quad N_1 = N_{11} + N_{21} + N_{31}, \quad N_2 = N_{22} + N_{32} + N_{42}, \quad \text{and} \quad K = N_{10} + 2N_{20} + N_{21} + 2N_{31} + N_{32} + 2N_{42}.
\]

With these notations, the maximum likelihood estimators for the allele frequency \( p \), the GRRs \( \Psi_1 \) and \( \Psi_2 \) and the plug-in estimator for AR are

\[
\hat{p} = \frac{K}{2n}, \quad \hat{\Psi}_1 = \frac{N_1}{N_0} \left( \frac{1 - \hat{p}}{2\hat{p}} \right), \quad \hat{\Psi}_2 = \left( \frac{N_2}{N_0} \right)^2 \left( \frac{1 - \hat{p}}{\hat{p}} \right)^2, \quad \hat{AR} = 1 - \frac{4nN_0}{(2n - K)^2}.
\]

Not surprisingly, the maximum likelihood estimator for \( p \) is the relative frequency of the D allele among the non-transmitted alleles. As already remarked by Knapp et al. (1995), the estimator \( \hat{\Psi}_1 \) can also be derived from the equation

\[
\frac{\text{Pr}(\text{Dd}|\text{affected})}{\text{Pr}(\text{dd}|\text{affected})} = \frac{\text{Pr}(\text{affected|Dd})\text{Pr}(\text{Dd})}{\text{Pr}(\text{affected|dd})\text{Pr}(\text{dd})},
\]

which is a direct consequence of the definition of conditional probabilities and transforms into

\[
\hat{\Psi}_1 = \frac{\text{Pr}(\text{Dd}|\text{affected})}{\text{Pr}(\text{dd}|\text{affected})} \cdot \frac{\text{Pr}(\text{Dd})}{\text{Pr}(\text{dd})}.
\]

Plugging-in the observed genotype frequencies in the sampled cases into the first factor and the allele frequencies estimated from the non-transmitted alleles into the second factor of this product yields the maximum likelihood estimator \( \hat{\Psi}_1 \). The estimator \( \hat{\Psi}_2 \) has an analogous representation.

Let

\[
\text{Logit}(x) = \ln \left( \frac{x}{1 - x} \right) \quad \text{and} \quad \text{InverseLogit}(y) = \frac{\text{exp}(y)}{1 + \text{exp}(y)}
\]

denote the Logit function and its inverse. Furthermore, let \( A = 1 + (2\hat{\Psi}_1 - 1)\hat{\rho} \) and \( B = 1 + (\hat{\Psi}_1 - 1)\hat{\rho}^2 + \hat{\rho} \) denote abbreviations. Approximate confidence intervals for a coverage probability of 1-\( \alpha \) for p, \( \Psi_1 \), \( \Psi_2 \) and AR are given by
3 Mathematical background

Table 3.1: All possible familial genotypes (family type) with observed frequencies and probabilities conditional upon the affection status of offspring; $N_{ij}$ is the number of trios with the corresponding familial genotype and $R = \Psi_2p^2 + \Psi_12p(1 - p) + (1 - p)^2$ (adapted from Scherag et al. (2002) with permission of S. Karger AG, Basel).

<table>
<thead>
<tr>
<th>family type</th>
<th>number observed</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>dd dd dd</td>
<td>$N_{00}$</td>
<td>$(1 - p)^2/R$</td>
</tr>
<tr>
<td>dd Dd dd</td>
<td>$N_{10}$</td>
<td>$2p(1 - p)^3/R$</td>
</tr>
<tr>
<td>dd Dd Dd</td>
<td>$N_{11}$</td>
<td>$\Psi_12p(1 - p)^3/R$</td>
</tr>
<tr>
<td>Dd Dd Dd</td>
<td>$N_{20}$</td>
<td>$p^2(1 - p)^2/R$</td>
</tr>
<tr>
<td>Dd Dd Dd</td>
<td>$N_{21\text{het}}$</td>
<td>$\Psi_14p^2(1 - p)^2/R$</td>
</tr>
<tr>
<td>Dd Dd DD</td>
<td>$N_{22}$</td>
<td>$\Psi_2p^2(1 - p)^2/R$</td>
</tr>
<tr>
<td>Dd DD Dd</td>
<td>$N_{31}$</td>
<td>$\Psi_12p^3(1 - p)/R$</td>
</tr>
<tr>
<td>Dd DD DD</td>
<td>$N_{32}$</td>
<td>$\Psi_2p^3(1 - p)/R$</td>
</tr>
<tr>
<td>DD DD DD</td>
<td>$N_{42}$</td>
<td>$\Psi_2p^4/R$</td>
</tr>
</tbody>
</table>

InverseLogit \left( \text{Logit}(\hat{p}) \pm \frac{z_1}{2} \sqrt{\frac{1}{2np(1 - \hat{p})}} \right),

\hat{\Psi}_1 \cdot \exp \left( \pm \frac{z_1}{2} \sqrt{\frac{\hat{\Psi}_2A\hat{p}^2 + (\hat{\Psi}_1 + A^2)(1 - \hat{p})}{\hat{\Psi}_12n\hat{p}(1 - \hat{p})^2}} \right),

\hat{\Psi}_2 \cdot \exp \left( \pm \frac{z_1}{2} \sqrt{\frac{\hat{\psi}_2(\hat{\psi}_2\hat{p}^4 + 2B\hat{p}(1 - \hat{p})) + A(1 - \hat{p})^3}{\hat{\psi}_2n\hat{p}^2(1 - \hat{p})^2}} \right),

and $\hat{AR} \pm \frac{z_1}{2} \sqrt{\frac{\hat{p}}{n(1 - \hat{p})^2} \cdot \frac{(\hat{\psi}_1 + 1)2(1 - \hat{p}) + \hat{\psi}_2\hat{p}}{(\hat{\psi}_2\hat{p}^2 + (1 - \hat{p})A)^2}}.$

The latter square root is equal to

$$\sqrt{\frac{1 - \hat{AR}}{n(1 - \hat{p})^2} \cdot \left[ \hat{AR} + (1 - \hat{AR})(\hat{p} + 3\hat{p}(1 - \hat{p})) \right]}.$$
Details about the derivation of the formulae are provided in Scherag et al. (2002).

In the case of small allele frequencies when normal approximations are not precise enough, exact Clopper-Pearson confidence intervals (Clopper and Pearson, 1934) for the allele frequency \( p \) can be obtained using the fact that the random variable \( K \) defined above (\( K \) = number of \( D \) alleles among the non-transmitted alleles) follows a binomial distribution with parameters \( 2n \) and \( p \). The resulting lower and upper confidence limits are

\[
p = \frac{K \cdot F_{2K,4n-2K+2,\alpha/2}}{2n - K + 1 + K \cdot F_{2K,4n-2K+2,\alpha/2}},
\]

and

\[
\bar{p} = \frac{(K + 1) \cdot F_{2K+2,4n-2K+2,1-\alpha/2}}{2n - K + (K + 1) \cdot F_{2K+2,4n-2K+2,1-\alpha/2}},
\]

where \( F_{k1,k2,\gamma} \) denotes the \( \gamma \)-quantile of the central \( F \)-distribution with \((k1,k2)\) df.

The point and the interval estimates described up to now depend on the Hardy-Weinberg equilibrium and will be called HWE confidence intervals. For the GRRs \( \hat{\Psi}_1 \) and \( \hat{\Psi}_2 \), point estimates and confidence intervals, which do not assume HWE, can be obtained by conditioning upon the parental genotypes (Schaid and Sommer, 1993). Unfortunately, since there is no explicit solution of the likelihood equation, numerical methods are required. The CPG likelihood and the method for the calculation of asymptotic CPG confidence intervals is described in Scherag et al. (2002). The CPG estimates and confidence intervals can also be obtained from a conditional logistic regression model (Cordell and Clayton, 2002). For this purpose a computer program written in C++ is provided in the Appendix.

When the population frequency of the risk allele is low and the associated mutation has a strong dominant effect, the number of homozygous allele carriers in the sample will be small or zero, and research interest may focus on the heterozygous relative risk. Exact confidence intervals for the heterozygous GRR \( \Psi_1 \) can be obtained from the conditional distribution of \( N_{11} \), given the sum \( \omega = N_{10} + N_{11} \), which is a binomial distribution with parameters \( \omega \) and \( \vartheta = \Psi_1/(1 + \Psi_1) \). From this binomial distribution, confidence limits
3 Mathematical background

for θ are obtained and then transformed into confidence limits for \( \Psi_1 \) by \( \Psi_1 = \theta/(1-\theta) \). The resulting estimator for \( \Psi_1 \) is \( \bar{\Psi}_1 = N_{11}/N_{10} \) with exact upper and lower confidence limits

\[
\Psi_1 = \frac{N_{11}}{N_{10} + 1} \cdot F_{2N_{11},2N_{10}+2,a/2}, \quad \text{and} \quad \bar{\Psi}_1 = \frac{N_{11} + 1}{N_{10}} \cdot F_{2N_{11}+2,2N_{10},1-a/2}.
\]

Note, that \( N_{11} \) and \( N_{10} \) denote the number of trios in which only one of the 4 parental alleles is a copy of \( D \) and in which the \( D \) allele is transmitted to the offspring and is not transmitted to the offspring, respectively. These estimates do not depend on the HWE assumption, since they are conditioned upon the parental genotype including only families with exactly one copy of \( D \) among the parental alleles. Of course, these confidence intervals are not fully efficient since information from other trios is ignored. This loss of efficiency will be small when the number of trios with offspring homozygous for the risk allele is small, as in the second example of the real data application described below.

Simulation study

Monte Carlo simulations were performed to study the small to moderate sample behavior of the asymptotic HWE confidence intervals. Trios were sampled from the multinomial distribution defined by the nine probability values given in the last column of Table 3.1. Simulations were performed under recessive, dominant and multiplicative genetic models defined by \( (\Psi_1, \Psi_2) = (1, 1.25), (1, 5), (1.25, 1.25), (5, 5), (1.25, 1.25^2), (5, 5^2), \) and \( (1, 1) \), which means no linkage and no association. For each model, allele frequencies \( p = 0.05, 0.1, 0.3, 0.5, 0.7, 0.9 \) and 0.95 were used. 10,000 simulations were run for all of these combinations of parameter values. Simulation results leading to a vanishing denominator in a formula were omitted from the calculation of the coverage probability. For the parameters \( \Psi_1 \) and \( \Psi_2 \), the simulation results confirmed the intuitive hypothesis that the validity of the confidence intervals should depend on the expected cell frequencies in Table 3.1, more specifically on the values \( e_{\Psi_1}(n) = \min\{E_n(N_0), E_n(N_1), E_n(K), n-E_n(K)\} \), \( e_{\Psi_2}(n) = \min\{E_n(N_0), E_n(N_2), E_n(K), n-E_n(K)\} \) and
$e_{AR}(n) = \min\{E_n(N_0), E_n(K), n - E_n(K)\}$, where $E_n(\cdot)$ denotes the expected value under a sample size of $n$. For given values of $p$, $\Psi_1$, $\Psi_2$, and $n$, the corresponding $e$-values can easily be calculated from Table 3.1 by multiplying all probability values by $n$ and summing up the corresponding cells. In Table 3.2, the numbers $n$ of trios required to achieve $e_{\Psi_1}(n) = 1$, $e_{\Psi_2}(n) = 1$, and $e_{AR}(n) = 1$ are tabulated for genetic models defined by different values of $\Psi_1$, $\Psi_2$ and $p$. Since $e(n) = n \cdot e(1)$, the $n$ required for any value $e > 1$ can be obtained by multiplying the tabulated value by $e$. In the simulations, for every parameter configuration $(\Psi_1, \Psi_2, p)$ sample sizes $n$ were ran up to $e_{\Psi_1}(n) = 15$, $e_{\Psi_2}(n) = 15$, and $e_{AR}(n) = 15$, but with a maximum of $n = 30,000$ for sake of practicability. The results (see Table 3.3) indicate that the nominal 95% confidence intervals for $\Psi_1$ and $\Psi_2$ hold the coverage probability or tend to be slightly conservative when $e_{\Psi_1}$ and $e_{\Psi_2}$ $\geq 5$, respectively. Asymptotic confidence intervals for $AR$ were found to be less reliable. For low allele frequencies (e.g., 0.05 and 0.1), the confidence intervals for $AR$ were found to be conservative or to nearly hold the coverage probability when $e_{AR} > 10$. However, for allele frequencies $\geq 0.5$, the confidence intervals for $AR$ were found to be moderately anti-conservative even for $e_{AR} = 15$. Monte Carlo results for the allele frequency $p$ are not reported, because fully efficient exact confidence intervals can always be calculated from the $F$-distribution as indicated above.

Sample sizes needed for asymptotic confidence intervals for $\Psi_2$ become impractical when the risk allele frequency $p$ is small and $\Psi_2$ itself is small to moderate. In this case, research interest will often focus on the heterozygous relative risk and Clopper-Pearson confidence intervals for $\Psi_1$ may be calculated without much loss of efficiency. Sample sizes for both $\Psi_1$ and $\Psi_2$ become impractical when the risk allele frequency is $> 0.7$ which is presumably not relevant for the investigation of complex diseases.

As a numerical example, consider the parameter values $p = 0.3$, $\Psi_1 = 1.25$ and $\Psi_2 = 1.25^2$ which are close to the estimates in the real data application given in the next section. As can be seen from the corresponding row of Table 3.2, sample sizes of 24, 82, and 24 trios will be sufficient to achieve $e_{\Psi_1}(n) = 10$, $e_{\Psi_2}(n) = 10$, and $e_{AR}(n) = 10$, and, hence, to obtain reliable confidence intervals under these parameter values.
Table 3.2: Required number of trios ($n$) to achieve $e_{\Psi_1}=1$, $e_{\Psi_2}=1$, and $e_{AR}=1$ for different genetic models. For $e \neq 1$, multiply the tabulated $n$-value by the desired $e$-value (adapted from Scherag et al. (2002) with permission of S. Karger AG, Basel).

<table>
<thead>
<tr>
<th>genetic model</th>
<th>$n$</th>
<th>$e_{\Psi_1}=1$</th>
<th>$e_{\Psi_2}=1$</th>
<th>$e_{AR}=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1 0.05</td>
<td>10.526</td>
<td>400.000</td>
<td>10.000</td>
<td></td>
</tr>
<tr>
<td>1 1 0.1</td>
<td>5.556</td>
<td>100.000</td>
<td>5.000</td>
<td></td>
</tr>
<tr>
<td>1 1 0.3</td>
<td>2.381</td>
<td>11.111</td>
<td>2.041</td>
<td></td>
</tr>
<tr>
<td>1 1 0.5</td>
<td>4.000</td>
<td>4.000</td>
<td>4.000</td>
<td></td>
</tr>
<tr>
<td>1 1 0.7</td>
<td>11.111</td>
<td>11.111</td>
<td>11.111</td>
<td></td>
</tr>
<tr>
<td>1 1 0.9</td>
<td>100.000</td>
<td>100.000</td>
<td>100.000</td>
<td></td>
</tr>
<tr>
<td>1 1 0.95</td>
<td>400.000</td>
<td>400.000</td>
<td>400.000</td>
<td></td>
</tr>
<tr>
<td>1 1.25 0.05</td>
<td>10.533</td>
<td>320.200</td>
<td>10.000</td>
<td></td>
</tr>
<tr>
<td>1 1.25 0.1</td>
<td>5.569</td>
<td>80.200</td>
<td>5.000</td>
<td></td>
</tr>
<tr>
<td>1 1.25 0.3</td>
<td>2.435</td>
<td>9.089</td>
<td>2.087</td>
<td></td>
</tr>
<tr>
<td>1 1.25 0.5</td>
<td>4.250</td>
<td>4.250</td>
<td>4.250</td>
<td></td>
</tr>
<tr>
<td>1 1.25 0.7</td>
<td>12.472</td>
<td>12.472</td>
<td>12.472</td>
<td></td>
</tr>
<tr>
<td>1 1.25 0.9</td>
<td>120.250</td>
<td>120.250</td>
<td>120.250</td>
<td></td>
</tr>
<tr>
<td>1 1.25 0.95</td>
<td>490.250</td>
<td>490.250</td>
<td>490.250</td>
<td></td>
</tr>
<tr>
<td>1 5 0.05</td>
<td>10.632</td>
<td>80.800</td>
<td>10.000</td>
<td></td>
</tr>
<tr>
<td>1 5 0.1</td>
<td>5.778</td>
<td>20.800</td>
<td>5.000</td>
<td></td>
</tr>
<tr>
<td>1 5 0.3</td>
<td>3.238</td>
<td>3.022</td>
<td>2.776</td>
<td></td>
</tr>
<tr>
<td>1 5 0.5</td>
<td>8.000</td>
<td>8.000</td>
<td>8.000</td>
<td></td>
</tr>
<tr>
<td>1 5 0.7</td>
<td>32.889</td>
<td>32.889</td>
<td>32.889</td>
<td></td>
</tr>
<tr>
<td>1 5 0.9</td>
<td>424.000</td>
<td>424.000</td>
<td>424.000</td>
<td></td>
</tr>
<tr>
<td>1 5 0.95</td>
<td>1844.000</td>
<td>1844.000</td>
<td>1844.000</td>
<td></td>
</tr>
<tr>
<td>1.25 1.25 0.05</td>
<td>10.000</td>
<td>327.800</td>
<td>10.000</td>
<td></td>
</tr>
<tr>
<td>1.25 1.25 0.1</td>
<td>5.000</td>
<td>83.800</td>
<td>5.000</td>
<td></td>
</tr>
<tr>
<td>1.25 1.25 0.3</td>
<td>2.301</td>
<td>10.022</td>
<td>2.301</td>
<td></td>
</tr>
<tr>
<td>1.25 1.25 0.5</td>
<td>4.750</td>
<td>4.750</td>
<td>4.750</td>
<td></td>
</tr>
<tr>
<td>1.25 1.25 0.7</td>
<td>13.639</td>
<td>13.639</td>
<td>13.639</td>
<td></td>
</tr>
<tr>
<td>1.25 1.25 0.9</td>
<td>124.750</td>
<td>124.750</td>
<td>124.750</td>
<td></td>
</tr>
<tr>
<td>1.25 1.25 0.95</td>
<td>499.750</td>
<td>499.750</td>
<td>499.750</td>
<td></td>
</tr>
<tr>
<td>5 5 0.05</td>
<td>10.000</td>
<td>111.200</td>
<td>10.000</td>
<td></td>
</tr>
<tr>
<td>5 5 0.1</td>
<td>5.000</td>
<td>35.200</td>
<td>5.000</td>
<td></td>
</tr>
<tr>
<td>5 5 0.3</td>
<td>6.204</td>
<td>6.756</td>
<td>6.204</td>
<td></td>
</tr>
<tr>
<td>5 5 0.5</td>
<td>16.000</td>
<td>16.000</td>
<td>16.000</td>
<td></td>
</tr>
<tr>
<td>5 5 0.7</td>
<td>51.556</td>
<td>51.556</td>
<td>51.556</td>
<td></td>
</tr>
<tr>
<td>5 5 0.9</td>
<td>496.000</td>
<td>496.000</td>
<td>496.000</td>
<td></td>
</tr>
<tr>
<td>5 5 0.95</td>
<td>1996.000</td>
<td>1996.000</td>
<td>1996.000</td>
<td></td>
</tr>
<tr>
<td>1.25 1.252 0.05</td>
<td>10.000</td>
<td>262.440</td>
<td>10.000</td>
<td></td>
</tr>
<tr>
<td>1.25 1.252 0.1</td>
<td>5.000</td>
<td>67.240</td>
<td>5.000</td>
<td></td>
</tr>
<tr>
<td>1.25 1.252 0.3</td>
<td>2.358</td>
<td>8.218</td>
<td>2.358</td>
<td></td>
</tr>
<tr>
<td>1.25 1.252 0.5</td>
<td>5.063</td>
<td>5.063</td>
<td>5.063</td>
<td></td>
</tr>
<tr>
<td>1.25 1.252 0.7</td>
<td>15.340</td>
<td>15.340</td>
<td>15.340</td>
<td></td>
</tr>
<tr>
<td>1.25 1.252 0.9</td>
<td>150.063</td>
<td>150.063</td>
<td>150.063</td>
<td></td>
</tr>
<tr>
<td>1.25 1.252 0.95</td>
<td>612.562</td>
<td>612.562</td>
<td>612.562</td>
<td></td>
</tr>
<tr>
<td>5 52 0.05</td>
<td>10.000</td>
<td>23.040</td>
<td>10.000</td>
<td></td>
</tr>
<tr>
<td>5 52 0.1</td>
<td>5.000</td>
<td>7.840</td>
<td>5.000</td>
<td></td>
</tr>
<tr>
<td>5 52 0.3</td>
<td>9.878</td>
<td>9.878</td>
<td>9.878</td>
<td></td>
</tr>
<tr>
<td>5 52 0.5</td>
<td>36.000</td>
<td>36.000</td>
<td>36.000</td>
<td></td>
</tr>
<tr>
<td>5 52 0.7</td>
<td>160.444</td>
<td>160.444</td>
<td>160.444</td>
<td></td>
</tr>
<tr>
<td>5 52 0.9</td>
<td>2116.000</td>
<td>2116.000</td>
<td>2116.000</td>
<td></td>
</tr>
<tr>
<td>5 52 0.95</td>
<td>9216.000</td>
<td>9216.000</td>
<td>9216.000</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.3: The median coverage probabilities and the range (in parenthesis) for the genetic models described in the text are tabulated. $e$ is the abbreviation for the quantity $e_{\Psi_1}$, $e_{\Psi_2}$, or $e_{AR}$ as defined in the text. For $AR$ separate columns for allele frequencies $p < 0.5$ or $p \geq 0.5$ are given (adapted from Scherag et al. (2002) with permission of S. Karger AG, Basel).

<table>
<thead>
<tr>
<th>$e$</th>
<th>for $\Psi_1$</th>
<th>for $\Psi_2$</th>
<th>for $AR$ ($p &lt; 0.5$)</th>
<th>for $AR$ ($p \geq 0.5$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.939</td>
<td>0.954</td>
<td>0.997</td>
<td>0.673</td>
</tr>
<tr>
<td></td>
<td>(0.914..1.000)</td>
<td>(0.913..0.998)</td>
<td>(0.629..1.000)</td>
<td>(0.581..0.997)</td>
</tr>
<tr>
<td>5</td>
<td>0.966</td>
<td>0.967</td>
<td>0.949</td>
<td>0.895</td>
</tr>
<tr>
<td></td>
<td>(0.960..0.979)</td>
<td>(0.951..0.975)</td>
<td>(0.909..0.963)</td>
<td>(0.880..0.913)</td>
</tr>
<tr>
<td>10</td>
<td>0.954</td>
<td>0.953</td>
<td>0.952</td>
<td>0.925</td>
</tr>
<tr>
<td></td>
<td>(0.949..0.961)</td>
<td>(0.947..0.962)</td>
<td>(0.932..0.956)</td>
<td>(0.920..0.933)</td>
</tr>
<tr>
<td>15</td>
<td>0.952</td>
<td>0.952</td>
<td>0.951</td>
<td>0.934</td>
</tr>
<tr>
<td></td>
<td>(0.949..0.959)</td>
<td>(0.947..0.960)</td>
<td>(0.940..0.955)</td>
<td>(0.931..0.941)</td>
</tr>
</tbody>
</table>

Real data application

The following examples contain real data, but will be presented here for numerical illustration only. Two different loci in two candidate genes were examined for association with the same phenotype - extreme early onset childhood and adolescent obesity. The observed familial genotypes are presented in Table 3.4. The first locus is a silent (C/T) SNP (rs133073) localized in the exon of candidate gene No. 1 - the melanin-concentrating hormone receptor 1 gene ($MCHRI$) on the human chromosome 22. The genetic results were published in Wermter et al. (2005). In candidate gene No. 2, a mutation screen was performed and different frameshift, nonsense and missense mutations were detected and grouped into risk and non-risk alleles according to the result of functional essays in transfected cells according to a recommendation by Hirschhorn and Altshuler (2002).

The resulting estimates and 95% confidence intervals (CI) for gene No. 1 are $\hat{p} = 0.329$ (95%CI 0.301...0.358), $\hat{\Psi}_1 = 1.458$ (95%CI 1.161...1.831), $\hat{\Psi}_2 = 1.855$ (95%CI 1.287...2.673) and $\hat{AR} = 0.228$ (95%CI 0.116...0.339), respectively. The fact that the
confidence intervals for $\Psi_1$ and $\Psi_2$ exclude 1 implies statistically significant evidence for association at a 0.05 level. The later conclusion would have also resulted from the TDT test statistic $T_{TDT} = 3.56$ (see chapter 2) with an asymptotic two-sided p-value of 0.00037 for the observed 286 transmission and 207 non-transmission of the risk allele C. Beyond the statistical proof of association, however, the confidence limits show that both homozygous and heterozygous carriers of risk alleles have an increase risk of disease, and they give a quantitative measure of the risk increments, which are small to moderate. According to the point estimates, the risk is higher for homozygous than for heterozygous carriers of the risk allele. Furthermore, the estimate for $AR$ indicates that the mutation associated with this SNP contributes to the development of the disease in a fairly large part (about 20%) of the diseased cases in the population from which the trios were sampled. The simulation study shows that the asymptotic confidence intervals are completely reliable under the parameter configuration and the sample size in the present example. In addition, CPG point estimates and confidence limits were calculated by the numerical method described in the Appendix. The estimates for gene No. 1 and the 95% CIs for $\Psi_1$ and $\Psi_2$ are $1.426$ (95%CI 1.120...1.816) and $1.873$ (95%CI 1.292...2.717), respectively, and thus do not substantially differ from the HWE estimates.

Due to the small cell frequencies, asymptotic methods are not appropriate for the second gene in Table 3.4. There are 11 transmissions and 3 non-transmissions, which yields a p-value of 0.029 in the exact one-sided McNemar test. Of the 1,040 non-transmitted alleles, three were a risk allele D, resulting in an estimate for the allele frequency $p$ of $3/1040 = 0.29\%$ with a Clopper-Pearson 95% CI of 0.0595 to 0.841%. There were 12 trios with exactly one risk allele in the parents, which was transmitted to 9 and not transmitted to 3 children. The conditional estimate for the heterozygous GRR obtained from these 12 trios is $9/3 = 3$, with an exact 95% CI of 0.749 to 17.228. There was only one homozygous ("compound heterozygous") carrier of functionally relevant mutations, so that the GRR for homozygous carriers cannot be estimated from these data.
Table 3.4: Observed frequencies of genotypes classified by family type in two candidate genes in a maximum of 529 trios (due to missing genotypes 9 trios were excluded for gene No. 2; in Wermter et al. (2005) 4 additional families were excluded due to missing genotypes at the other explored loci; (adapted from Scherag et al. (2002) with permission of S. Karger AG, Basel).

<table>
<thead>
<tr>
<th>family type</th>
<th>gene No.1</th>
<th>gene No.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>parental</td>
<td>genotype of affected offspring</td>
<td></td>
</tr>
<tr>
<td>dd</td>
<td>dd</td>
<td>78</td>
</tr>
<tr>
<td>dd</td>
<td>Dd</td>
<td>90</td>
</tr>
<tr>
<td>dd</td>
<td>Dd</td>
<td>121</td>
</tr>
<tr>
<td>Dd</td>
<td>Dd</td>
<td>16</td>
</tr>
<tr>
<td>Dd</td>
<td>Dd</td>
<td>57</td>
</tr>
<tr>
<td>dd</td>
<td>DD</td>
<td>57</td>
</tr>
<tr>
<td>Dd</td>
<td>Dd</td>
<td>36</td>
</tr>
<tr>
<td>Dd</td>
<td>DD</td>
<td>28</td>
</tr>
<tr>
<td>Dd</td>
<td>DD</td>
<td>36</td>
</tr>
<tr>
<td>DD</td>
<td>DD</td>
<td>10</td>
</tr>
<tr>
<td>sum</td>
<td></td>
<td>529</td>
</tr>
</tbody>
</table>

Discussion

The likelihood theory for the case parents design has been extensively studied by various authors (Schaid, 1999; Knapp et al., 1995; Schaid and Sommer, 1994, 1993). The results of this work are a supplement to their work. Explicit formulae for easy calculation of confidence intervals for the GRRs, the allele frequencies, and the attributable risk were developed. Their validity was investigated by Monte Carlo simulations and their application was demonstrated by numerical examples. The formulae for HWE confidence intervals are based on the unconditional likelihood model, which assumes HWE. The confidence intervals may be biased due to deviations from HWE. Thus, it is recommended to compare HWE and CPG estimates if possible.

One should also note that the presented methods are only applicable to data sets of case parent trios which have been ascertained for the affected offspring. These are trios ascertained for the presence of an affected child, where a biallelic maker locus identical
3 Mathematical background

to the disease locus, i.e., recombination rate $\theta = 0$ and complete linkage disequilibrium is assumed. The computer code provided in the Appendix also allows for obtaining estimates given a situation of more than one affected child.

In general, the estimation of genetic effect sizes provides more information than a mere statistical test such as the TDT (Spielman et al., 1993). GRRs can be thought of as an individual measure of the genetic risk of mutation carriers to develop the disease, while the attributable risk is a population-based measure of the epidemiological relevance of the mutation. However, for the interpretation of GRR estimates the simplification underlying the statistical model should be taken into account. It is assumed that the investigated candidate gene contributes in a multiplicative way to the risk of disease, and that the risk ratio does not depend on the genetic variants of the individual in other relevant genes. Clearly, in most complex diseases the contribution of the specific investigated gene will depend on the baseline risk in the sense that it will be smaller in individuals loaded with additional predisposing genetic factors. Thus, the estimates will of course depend on the recruitment in the special study. For example, high rates of bilineal parental occurrence of the phenotype in the sampled trios may reduce the GRR estimate at the special investigated locus. This remains true even when CPG estimates are used, which are unaffected by assortative mating only within the statistical model of multiplicative risks over different loci.

Finally, one should carefully consider the usefulness of the formulae in some situations. Since the number of sampled trios required for reliable estimation of $\Psi_1$, $\Psi_2$ and $AR$ increases for very high or very low allele frequencies and also increases with increased true values of GRRs, the formulae might not be applied to a constellation in which very rare expected allele frequencies and very high GRRs are investigated. The $e$-values proposed in the section on Monte Carlo simulations may be a useful guidance for the application of the formulae.
4 Data adaptive interim modification of sample sizes for candidate gene association studies

4.1 Introduction

As already outlined in the previous chapter, sample size calculations and statistical power considerations crucially depend on the true effect sizes within an investigated population. Overestimation of these effect sizes at the design stage will result in an underpowered study - underestimation can lead to excess use of resources due to extremely large sample sizes.

In this chapter, a method for design adaptations will be used to allow for sample size recalculations depending on interim estimates of effect sizes (GRRs) and allele frequencies as described in the previous chapter. The underlying general principle of the method (Müller and Schäfer, 2004, 2001) was described in chapter 3. Here its applicability to genetic research will be demonstrated by adapting it to the TDT (Spielman et al., 1993) for the case parent trio design in candidate gene studies. Parts of this chapter have been published in Scherag et al. (2003).

The adaptive procedure will be combined with classical group sequential plans which were introduced for clinical trials by Pocock (1977), O’Brien and Fleming (1979), and were subsequently refined by other authors (reviewed in Jennison and Turnbull, 1999). As a consequence, such designs will include both the option of early rejection of the null hypothesis ($H_0$) of no linkage and no association during the course of the study and the option for design adaptation and sample size recalculations when $H_0$ cannot be rejected in the interim analysis. To this end, a sequential version of the TDT is developed
which can be approximated by a Brownian motion process (see chapter 3). The limiting Brownian process will be shown to have non-zero drift and a variance parameter different from 1, under the alternative hypothesis. Thus, a two parameter stochastic process is needed for adequate sequential modeling of the TDT. In addition, to demonstrate the wider range of possible applications, a real data example for a case-control association study is provided.

Note that stopping the study after an interim analysis and accepting $H_0$ (futility stops) will not be dealt with in this chapter as such it will less likely apply to a candidate gene investigation. On a genomewide level, it is a necessity for marker reduction and will thus be implemented in the procedure described in chapter 5.

4.2 Method

4.2.1 Fixed sample, group sequential and data adaptive plans

In the following, three methods of sampling will be distinguished - fixed sample designs (FS), classical group sequential plans (GS) and an adaptive sequential procedure (AS). The term fixed sample designs refers to studies for which a pre-specified number of subjects is sampled without permitting for statistical testing ahead of time - exactly one statistical test will be performed for the whole data. Classical group sequential plans allow for stopping an experiment early with rejection of $H_0$ when there is sufficient statistical evidence at an interim analysis (Jennison and Turnbull, 1999; Lan and DeMets, 1983; O’Brien and Fleming, 1979; Pocock, 1977). However, when this is not the case and the study is continued, classical GS plans do not allow for modifying the further sample size of the study. Adaptive procedures, introduced in chapter 3, offer the possibility to re-determine the sample size and to modify the study design depending on interim effect size estimates (Chang, 2007; Müller and Schäfer, 2004, 2001; Lehmacher and Wassmer, 1999; Cui et al., 1999; Fisher, 1998; Proschan and Hunsberger, 1995; Bauer and Köhne, 1994). Even though some of the adaptive designs will be used for comparisons in the simulation studies below, the method for design adaptations by Müller and Schäfer (2004, 2001) will be of primary interest in this and the subsequent chapter. It can be combined
with any group sequential plan, allows for data-dependent modifications of the sample size, of the number and time points of interim analyses, and of the group sequential boundaries. These design changes can be made at any time during the study. This will result in a procedure that combines interim hypothesis testing and early rejection of $H_0$ with the option of sample size recalculation depending on the observed interim effect sizes. Thus, it will be called adaptive sequential design (AS).

In more detail, the AS works as follows: A two-stage conventional GS design is selected and an initial sample size calculation is done for this GS design. For the appropriate number of trios in the first stage, the interim analysis is performed according to the interim boundaries of the selected GS design. If $H_0$ (no linkage and no association) cannot be rejected, the sample size for the second stage is recalculated based on the estimates of the effect size obtained at the interim analysis. Without adjustment of the final decision boundaries, such data-driven sample size adaptations may lead to an inflation of the type I error risk (Armitage et al., 1969). Using the CRP principle (Müller and Schäfer, 2004), control of the overall type I error is provided. The CRP principle is very flexible and allows for any kind of sample size revision at any time during the study; here a pre-specified function will be used to calculate the residual sample size depending on the interim effects size estimates and the interim allele frequency estimates.

### 4.2.2 Approximation of the TDT by a Brownian motion process - group sequential TDT

As before, suppose a biallelic disease susceptibility locus with alleles $D$ and $d$, where again $D$ is the putative risk allele. The parameters of interest are the allele frequencies $p$ and $(1-p)$ of the alleles $D$ and $d$, respectively, and the GRRs $\Psi_1$ and $\Psi_2$ for heterozygous and homozygous carriers of allele $D$ as defined in chapter 3. To test for genetic association and linkage in trios, the TDT statistic (Spielman et al., 1993) will be used. Here, the investigated marker represents the disease locus itself, so that the recombination fraction $\theta = 0$ and there is complete LD. To introduce a sequential version of the TDT, let $t = n_1/n$, $t \in [0,1]$ denote the proportion of $n_1$ trios relative to the total planned
sample size \( n \). For a given value of this information time parameter \( t \), let \( N^T(t) \) be the observed number of heterozygous parents transmitting the putative risk allele \( D \) and let \( N^{NT}(t) \) be the number of heterozygous parents not transmitting \( D \) among the first \( n_1 \) trios. The sequential version of the TDT is given by

\[
T_{TDT}(t) = \sqrt{t} \cdot \frac{U(t)}{\sqrt{\text{Var}(U(t))}}
\]

with \( U(t) = N^T(t) - N^{NT}(t) \) and \( \text{Var}(U(t)) = N^T(t) + N^{NT}(t) \).

Asymptotically, \( T_{TDT}(t) \) follows a Brownian motion process with a drift parameter

\[
\sqrt{n} \cdot \mu(p, \Psi_1, \Psi_2) = -\sqrt{n} \cdot \frac{\sqrt{2}[1 - \Psi_1 + p(2\Psi_1 - \Psi_2 - 1)]\sqrt{(p-1)p[1+\Psi_1+p(\Psi_2-1)]} - \sqrt{(p-1)p[1+\Psi_1+p(\Psi_2-1)]-p^2\Psi_2}}{1 + \Psi_1 + p(\Psi_2 - 1)},
\]

and a variance parameter

\[
\sigma^2(p, \Psi_1, \Psi_2) = \frac{1}{4} \left( p - \frac{4(p-1)(2p-1)[1 + p(\Psi_1 - 1)]^2(1 + \Psi_1)}{[1 + \Psi_1 + p(\Psi_2 - 1)]^3} - \frac{4[1 + p(\Psi_1 - 1)][3 + \Psi_1 + 3p\Psi_1 - p^2(3 + \Psi_1)]}{[1 + \Psi_1 + p(\Psi_2 - 1)]^2} - \frac{15 - \Psi_1 - p\{47 - 13\Psi_1 - 4p[7 + p + \Psi_1(p - 7)]\}}{(2p - 1)[1 + \Psi_1 + p(\Psi_2 + 1)]} - \frac{2(p - 1)[1 + p(\Psi_1 + 1)]}{(2p - 1)\{p - 1[1 + p(2\Psi_1 - 1)] - p^2\Psi_2\}} \right)
\]

(for details review chapter 3). The marginal distribution for a fixed \( t \) is given by

\[
T_{TDT}(t) \sim \mathcal{N}(\sqrt{n} \cdot \mu(p, \Psi_1, \Psi_2) \cdot t, \sigma^2(p, \Psi_1, \Psi_2) \cdot t).
\]

where under \( H_0 \), both \( \Psi_1 \) and \( \Psi_2 \) equal 1 which leads to \( \mu(p, \Psi_1, \Psi_2) = 0 \) and \( \sigma^2(p, \Psi_1, \Psi_2) = 1 \), respectively. Thus, under \( H_0 \), \( \sqrt{t} \cdot T_{TDT}(t) \) follows a standard Brownian motion and conventional group sequential boundaries for a standard Brownian motion can be applied. In the first simulation study and in the application, a two-stage Pocock design (Pocock, 1977) with \( \alpha = 0.05 \) (two-sided) and \( t_{\text{interim}} = 0.5 \) was used which means that one interim analysis will be performed after half of the initially planned total sample size. The critical values for \( T_{TDT}(0.5) \) and \( T_{TDT}(1) \) are \( \pm c_1 = \pm \sqrt{0.5} \cdot 2.178 \) and \( \pm c_2 = \)
Sample size modifications for candidate gene studies

$H_0$ is rejected if $T_{TDT}(0.5)$ or $T_{TDT}(1)$ falls outside these limits. A second simulation study deals with the consequences of alternative choices of $t_{interim}$.

Next, suppose that the investigator wants to achieve an overall power of $1 - \beta$ for pre-specified parameter values $p^*$, $\Psi_1^*$, and $\Psi_2^*$. For the FS design the required number of trios is

$$n_{FS} = \frac{\left[z_{1-\alpha/2} + z_{1-\beta} \cdot \sigma(p^*, \Psi_1^*, \Psi_2^*)\right]^2}{\mu^2(p^*, \Psi_1^*, \Psi_2^*)},$$

where $z_\gamma$ denotes the $\gamma$-quantile of a standard normal distribution and $\alpha$ is the significance level.

For GS designs, sample size determination requires the calculation of the total transition probability of the group sequential boundaries defined above, under the Brownian motion process with drift $\delta = \sqrt{n} \cdot \mu(p^*, \Psi_1^*, \Psi_2^*)$ and variance parameter $\sigma^2 = \sigma^2(p^*, \Psi_1^*, \Psi_2^*)$ as defined in the chapter 3. The numerical integration can either be done by the function \text{SEQ} as implemented in SAS/IML (2001), by the R package (2007) \texttt{mvtnorm}, or by the C++ program provided in the Appendix. The appropriate maximum sample size $n_{GS}$ can be found by a manual search over different values of $n_{GS}$ or by a bisection algorithm. In the case of the two-stage Pocock design (Pocock, 1977) a good rule of thumb is a 10% increment over the sample sizes for the FS design.

4.2.3 Interim analysis and sample size adaptation

If the interim boundaries are not crossed, the remaining part of the study can be redesigned using the CRP principle. The first step is to calculate the conditional probability that $H_0$ will be rejected at the final analysis, given the interim value of the TDT test statistic $T_{TDT}(t_{interim})$, if $H_0$ is actually true: $Pr_{H_0}(\text{reject } H_0 | \text{interim value})$. The CRP can be approximately determined via the Brownian motion approximation of the TDT. For the special case of a two-stage GS design considered in the present chapter or a FS design, CRPs are given by

$$CRP_{lower} = \Phi \left( \frac{-c_2 - T_{TDT}(t_{interim})}{\sqrt{1 - t_{interim}}} \right),$$

and

$$\text{and } CRP_{upper} = 1 - \Phi \left( \frac{c_2 - T_{TDT}(t_{interim})}{\sqrt{1 - t_{interim}}} \right),$$

(4.2)
where $\Phi(\cdot)$ denotes the cumulative distribution of a standard normal variable, $\pm c_2$ are the critical values for $T_{TDT}(1)$ at the final analysis and $t_{\text{interim}}$ denotes the proportion of trios collected up to the interim analysis relative to the initially planned sample size $n$. Thus, $1 - t_{\text{interim}}$ is the remaining proportion of trios to be collected after the interim analysis according to the initial plan. The CRP principle says that once the investigator has calculated $\text{CRP}_{\text{lower}}$ and $\text{CRP}_{\text{upper}}$, which guarantees the control of the overall type I error at the pre-specified significance level $\alpha$, she or he may select any design for the rest of the study, with the only constraint that the new design has to hold the upper and lower type I error risks of $\text{CRP}_{\text{lower}}$ and $\text{CRP}_{\text{upper}}$. Further statistical testing is then solely based on the new data, since information from the already collected sample is “preserved” in the CRPs. In the case of observing a large interim effect for which one of the CRPs will be very small, the investigator may decide in favor of a one-sided study continuation with only the larger CRP serving as the new significance level.

The general principle underlying the idea of calculating CRPs can also be applied to GS designs with more than two stages with or without futility stops or to designs with multiple interim analyses. Again, these calculations will require the numerical integration methods mentioned above.

Assume that the initially planned sample size was $n$ and consists of $n = n_1 + n_2$ in the case of the two-stage GS design, where $n_1$ and $n_2$ are the sample sizes of the corresponding stages. For the remainder of this chapter, suppose that an FS design will be used for the re-designed second stage of the study. Equation 4.1 can then be used for sample size adaptation. One may plug-in interim estimates $\hat{\mu}$, $\hat{\Psi}_1$, and $\hat{\Psi}_2$ based on the $n_1$ trios which were sampled up to the interim analysis. Remember that suitable estimators for $p$, $\Psi_1$, and $\Psi_2$ were described in chapter 3.

The resulting plug-in estimates for $\mu$ and $\sigma$ will be denoted by $\hat{\mu}$ and $\hat{\sigma}$, i.e., $\hat{\mu} = \mu(\hat{\mu}, \hat{\Psi}_1, \hat{\Psi}_2)$ and $\hat{\sigma} = \sigma(\hat{\mu}, \hat{\Psi}_1, \hat{\Psi}_2)$. According to a proposal by Shen and Fisher (1999), the initial overall power $1 - \beta$ will be replaced by the conditional power $1 - \beta^*$, which is defined as the conditional probability that $H_0$ will be rejected at the final analysis given that it has not been rejected in the interim analysis and that $H_1$ for $\hat{\mu}$ and $\hat{\sigma}$ is actually
true (for details see Scherag et al., 2003).

Next, $\alpha$ will be substituted by $CRP = \max\{CRP_{\text{lower}}, CRP_{\text{lower}}\}$, which means that one believes in the direction of the interim estimate. Finally, for protection against instabilities in the interim estimates and against impractically large sample sizes, a truncation rule for the new sample size $n^*_2$ was implemented for the simulation studies. Thus, the final proposal is

$$n^*_2 = \begin{cases} \frac{[z_{1 - CRP}+z_{1 - \beta^*}]}{\bar{q}^2}^2 & \text{if } \frac{[z_{1 - CRP}+z_{1 - \beta^*}]}{\bar{q}^2}^2 < n_{\max} \\ n_2 & \text{if } \frac{[z_{1 - CRP}+z_{1 - \beta^*}]}{\bar{q}^2}^2 \geq n_{\max} \end{cases}$$

To begin with, the truncation value selected was $n_{\max} = 4n$ for the first and the second simulation study and the numerical example. The initial design, including the original critical values for testing, will be maintained if the recalculated sample size is larger than $n_{\max}$. Note that, for now, this proposal does not include a formal decision rule for the acceptance of $H_0$ at the interim analysis. This issue will be dealt with in chapter 5.

If the sample size is reassessed, a further number of $n^*_2$ trios has been collected and an ordinary TDT test is performed based solely on these $n^*_2$ trios. A conventional one-sided p-value for testing the remaining one-sided null hypothesis is calculated. This p-value is compared to $CRP$ and the null hypothesis $H_0$ of no linkage and no association can be rejected if the observed p-value $\leq CRP$.

### 4.3 Simulation studies

#### 4.3.1 Comparison of FS, GS, and AS designs

To evaluate some properties of the proposed AS design in comparison to the FS and GS designs, Monte Carlo simulations were performed. As similar results were observed for other genetic models and parameter values, the simulations described will be restricted to multiplicative ($\Psi_1 = \Psi, \Psi_2 = \Psi^2$) genetic models with $\Psi \in \{1, 1.3, 1.5, 1.7\}$, and allele frequencies $p$ of the risk allele $p \in \{0.05, 0.1, 0.3\}$. Trios were sampled from the multinominal distribution of ten familial genotypes as described in Table 1 of Schaid and Sommer (1993). The FS design, the two-stage Pocock GS design (Pocock, 1977), and
the AS design were compared with each other. For each true model parameter pair $p$ and $\Psi$ the behavior of all three designs was investigated for different pre-study estimates $p^*$ and $\Psi^*$ with $p^* = p \pm 0.03$ and $\Psi^* = \Psi \pm 0.2$. To evaluate $H_0$, i.e., $\Psi=1$, $n=1,000$ was assumed. For each parameter combination 100,000 replicates were evaluated. For both the GS and the AS design, $t_{\text{interim}} = 0.5$ was chosen. The Monte Carlo type I error rate and the Monte Carlo power as well as the average sample sizes ($n_{\text{average}}$) and the standard deviations of the sample sizes were obtained. In addition, Table 4.1-4.3 show the proportion of all replicates for which the initial GS sample size planning with $n_2$ was maintained (abbreviated as $% n_2$).

For all three designs the type I error rates range between 0.047 and 0.051. When the pre-study estimates are smaller than the true effects, e.g., $p^* = p - 0.03$ and $\Psi^* = \Psi - 0.2$, average sample sizes under the AS and the GS design are up to 45% lower than the sample size of the FS design. Furthermore, Table 4.1 shows that the estimated power for the AS design was always the same or larger than the power for the classical FS and GS designs. As expected, this gain in power leads to increased average sample sizes for the AS procedure compared to the average sample sizes for the GS design. When the pre-study estimates are larger than the true effects, the AS design will frequently produce sample size extensions with the consequence of a power gain compared to the FS or GS designs. The power gain ranges between 7 and 9% for $\Psi^* = \Psi + 0.2$ and between 4 and 7% for $p^* = p + 0.03$. Note that the gain in power can easily be controlled by the choice of $n_{\text{max}}$ which will be evaluated in more detail in the next section.

4.3.2 Comparison with adaptive designs

In order to extend the view on the proposed AS design, an additional comparison was done to investigate some of the adaptive designs introduced in chapter 3. In particular, the considered adaptive two-stage designs were Fisher’s combination rule (Fisher, 1932), the “variance spending”- or “self-designing”-method (Fisher, 1998), and a recent proposal by Chang (2007) for which the overall probability for a type I error is given by
Table 4.1: Comparison of the FS, GS, and AS designs (100,000 replicates): Results of the Monte Carlo simulation for multiplicative genetic models and the TDT. Displayed are the variables for sample size ($n$, $n_{avg}$, STD$_n$, % $n_2$) and the overall power. $\Psi$: true GRR parameter, $p$: true allele frequency of disease allele, for planning: $\Psi^*$: assumed GRR parameter, $p^*$: assumed disease allele frequency, $\alpha$ (two-sided) = 0.05, and $\beta$ = 0.2; italics indicate perfect matches of planning and truth (adapted from Scherag et al. (2003) with permission of S. Karger AG, Basel).

<table>
<thead>
<tr>
<th>$\Psi$</th>
<th>$p$</th>
<th>$\Psi^*$</th>
<th>$p^*$</th>
<th>FS design</th>
<th>GS design</th>
<th>AS design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>power</td>
<td>$n_{avg}$</td>
<td>STD$_n$</td>
<td>power</td>
</tr>
<tr>
<td>1.3</td>
<td>0.05</td>
<td>17,428</td>
<td>1.000</td>
<td>19,350</td>
<td>9,677</td>
<td>80.00</td>
</tr>
<tr>
<td>1.3</td>
<td>0.02</td>
<td>5,128</td>
<td>0.991</td>
<td>5,698</td>
<td>3,267</td>
<td>1,007.00</td>
</tr>
<tr>
<td>1.3</td>
<td>0.05</td>
<td>2,135</td>
<td>0.801</td>
<td>2,372</td>
<td>1,827</td>
<td>593.00</td>
</tr>
<tr>
<td>1.3</td>
<td>0.08</td>
<td>1,390</td>
<td>0.615</td>
<td>1,544</td>
<td>1,306</td>
<td>358.00</td>
</tr>
<tr>
<td>1.5</td>
<td>0.05</td>
<td>839</td>
<td>0.418</td>
<td>934</td>
<td>845</td>
<td>184.00</td>
</tr>
<tr>
<td>1.3</td>
<td>0.1</td>
<td>9,244</td>
<td>1.000</td>
<td>10,263</td>
<td>5,132</td>
<td>43.00</td>
</tr>
<tr>
<td>1.3</td>
<td>0.07</td>
<td>1,567</td>
<td>0.908</td>
<td>1,741</td>
<td>1,216</td>
<td>426.00</td>
</tr>
<tr>
<td>1.3</td>
<td>0.1</td>
<td>1,144</td>
<td>0.801</td>
<td>1,270</td>
<td>978</td>
<td>321.00</td>
</tr>
<tr>
<td>1.3</td>
<td>0.13</td>
<td>918</td>
<td>0.710</td>
<td>1,020</td>
<td>830</td>
<td>247.00</td>
</tr>
<tr>
<td>1.5</td>
<td>0.1</td>
<td>454</td>
<td>0.423</td>
<td>505</td>
<td>457</td>
<td>100.00</td>
</tr>
<tr>
<td>1.3</td>
<td>0.3</td>
<td>4,040</td>
<td>1.000</td>
<td>4,486</td>
<td>2,243</td>
<td>27.00</td>
</tr>
<tr>
<td>1.3</td>
<td>0.27</td>
<td>548</td>
<td>0.821</td>
<td>609</td>
<td>461</td>
<td>152.00</td>
</tr>
<tr>
<td>1.3</td>
<td>0.3</td>
<td>519</td>
<td>0.801</td>
<td>576</td>
<td>443</td>
<td>144.00</td>
</tr>
<tr>
<td>1.3</td>
<td>0.33</td>
<td>497</td>
<td>0.784</td>
<td>552</td>
<td>429</td>
<td>137.00</td>
</tr>
<tr>
<td>1.5</td>
<td>0.3</td>
<td>213</td>
<td>0.431</td>
<td>237</td>
<td>214</td>
<td>48.00</td>
</tr>
<tr>
<td>1.5</td>
<td>0.05</td>
<td>2,135</td>
<td>0.994</td>
<td>2,372</td>
<td>1,334</td>
<td>391.00</td>
</tr>
<tr>
<td>1.5</td>
<td>0.02</td>
<td>2,004</td>
<td>0.991</td>
<td>2,230</td>
<td>1,278</td>
<td>394.00</td>
</tr>
<tr>
<td>1.5</td>
<td>0.05</td>
<td>839</td>
<td>0.802</td>
<td>934</td>
<td>718</td>
<td>232.00</td>
</tr>
<tr>
<td>1.5</td>
<td>0.08</td>
<td>550</td>
<td>0.620</td>
<td>611</td>
<td>517</td>
<td>142.00</td>
</tr>
<tr>
<td>1.7</td>
<td>0.05</td>
<td>464</td>
<td>0.547</td>
<td>517</td>
<td>451</td>
<td>113.00</td>
</tr>
<tr>
<td>ψ</td>
<td>p</td>
<td>true values</td>
<td>initial plan</td>
<td>FS design</td>
<td>GS design</td>
<td>AS design</td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-------------</td>
<td>--------------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>n avg</td>
<td>STD</td>
<td>n</td>
<td>n avg</td>
</tr>
<tr>
<td>1.5</td>
<td>0.1</td>
<td>1.144</td>
<td>1.270</td>
<td>213</td>
<td>0.995</td>
<td>7.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>618</td>
<td>688</td>
<td>168</td>
<td>0.909</td>
<td>572</td>
</tr>
<tr>
<td>1.5</td>
<td>0.1</td>
<td>454</td>
<td>505</td>
<td>126</td>
<td>0.799</td>
<td>491</td>
</tr>
<tr>
<td>1.5</td>
<td>0.13</td>
<td>366</td>
<td>407</td>
<td>99</td>
<td>0.708</td>
<td>431</td>
</tr>
<tr>
<td>1.7</td>
<td>0.1</td>
<td>253</td>
<td>282</td>
<td>62</td>
<td>0.543</td>
<td>332</td>
</tr>
<tr>
<td>1.5</td>
<td>0.3</td>
<td>619</td>
<td>576</td>
<td>100</td>
<td>0.994</td>
<td>344</td>
</tr>
<tr>
<td>1.5</td>
<td>0.27</td>
<td>224</td>
<td>249</td>
<td>62</td>
<td>0.822</td>
<td>236</td>
</tr>
<tr>
<td>1.5</td>
<td>0.3</td>
<td>213</td>
<td>237</td>
<td>60</td>
<td>0.800</td>
<td>229</td>
</tr>
<tr>
<td>1.5</td>
<td>0.33</td>
<td>205</td>
<td>228</td>
<td>57</td>
<td>0.762</td>
<td>228</td>
</tr>
<tr>
<td>1.7</td>
<td>0.3</td>
<td>123</td>
<td>136</td>
<td>31</td>
<td>0.564</td>
<td>164</td>
</tr>
<tr>
<td>1.7</td>
<td>0.05</td>
<td>839</td>
<td>934</td>
<td>206</td>
<td>0.967</td>
<td>657</td>
</tr>
<tr>
<td>1.7</td>
<td>0.02</td>
<td>1,102</td>
<td>1,227</td>
<td>219</td>
<td>0.992</td>
<td>712</td>
</tr>
<tr>
<td>1.7</td>
<td>0.05</td>
<td>464</td>
<td>517</td>
<td>129</td>
<td>0.801</td>
<td>498</td>
</tr>
<tr>
<td>1.7</td>
<td>0.08</td>
<td>306</td>
<td>340</td>
<td>79</td>
<td>0.618</td>
<td>381</td>
</tr>
<tr>
<td>1.9</td>
<td>0.05</td>
<td>303</td>
<td>337</td>
<td>78</td>
<td>0.613</td>
<td>377</td>
</tr>
<tr>
<td>1.7</td>
<td>0.1</td>
<td>454</td>
<td>505</td>
<td>112</td>
<td>0.967</td>
<td>358</td>
</tr>
<tr>
<td>1.7</td>
<td>0.07</td>
<td>343</td>
<td>382</td>
<td>94</td>
<td>0.909</td>
<td>320</td>
</tr>
<tr>
<td>1.7</td>
<td>0.1</td>
<td>253</td>
<td>282</td>
<td>71</td>
<td>0.802</td>
<td>275</td>
</tr>
<tr>
<td>1.7</td>
<td>0.13</td>
<td>206</td>
<td>229</td>
<td>56</td>
<td>0.712</td>
<td>243</td>
</tr>
<tr>
<td>1.9</td>
<td>0.1</td>
<td>167</td>
<td>186</td>
<td>43</td>
<td>0.616</td>
<td>209</td>
</tr>
<tr>
<td>1.7</td>
<td>0.3</td>
<td>213</td>
<td>237</td>
<td>53</td>
<td>0.964</td>
<td>171</td>
</tr>
<tr>
<td>1.7</td>
<td>0.27</td>
<td>129</td>
<td>143</td>
<td>36</td>
<td>0.823</td>
<td>138</td>
</tr>
<tr>
<td>1.7</td>
<td>0.3</td>
<td>123</td>
<td>136</td>
<td>34</td>
<td>0.806</td>
<td>137</td>
</tr>
<tr>
<td>1.7</td>
<td>0.33</td>
<td>119</td>
<td>132</td>
<td>33</td>
<td>0.787</td>
<td>133</td>
</tr>
<tr>
<td>1.9</td>
<td>0.3</td>
<td>83</td>
<td>93</td>
<td>22</td>
<td>0.636</td>
<td>104</td>
</tr>
</tbody>
</table>
\[ \alpha = \alpha_1 + \frac{1}{2}(\alpha_2 - \alpha_1)^2. \]

A comparison of the AS design with the adaptive design by Lehmacher and Wassmer (1999) was omitted as it would lead to identical results within the applied Brownian motion process representation.

In either case, the true multiplicative genetic model was given by \( \Psi = 1.3 \) and \( p = 0.2 \) and the initial planning for a two-stage Pocock GS design (Pocock, 1977) again varied (see Table 4.2). For each design \( t_{interim} \in \{0.1, 0.25, 0.5, 0.75\} \) was used to investigate the impact of early, mid course or late interim evaluations as well. Moreover, different values of \( n_{max} \in \{1.5n, 2.5n, 4n\} \) for the sample size truncation rule (4.3) were explored.

As before, reassessed sample sizes larger than \( n_{max} \) resulted in maintaining the originally planned sample size \( n_2 \) of the GS design and, in case of the AS design, even in an adherence to the GS critical values.

To evaluate \( H_0 \) again \( n=1,000 \) and \( t_{interim} = 0.5 \) was explored. As expected for a one-sided test with \( \alpha(\text{one-sided}) = 0.025 \), all type I error rates ranged between 0.025 and 0.026.

The comparison of the designs revealed that both the AS design and Fisher’s combination rule (Fisher, 1932) outperformed the adaptive designs of Fisher (1998) and Chang (2007) by and large. While the AS design always had smaller average samples sizes compared to the “self-designing” adaptive design by Fisher (1998), its power was up to 13% larger, except for the case of very early interim analyses \( (t_{interim} = 0.1) \) for a properly powered or an underpowered study \( (p^* \in \{0.2,0.3\}) \). For this situation a slight inferiority of about 3% was observable for the AS design. The comparison to the adaptive design by Chang (2007) revealed that the AS design was always more powerful (up to 32%), especially for earlier interim analyses \( (t_{interim} \in \{0.1, 0.25\}) \). Note that the variables for sample size \( (n, n_{avg}, \text{STD}_n, \% n_2) \) are identical for the AS design and the design by Chang (2007) which is why an additional description was omitted in Table 4.2.

In contrast, the AS design and Fisher’s combination rule (Fisher, 1932) were less
Table 4.2: Comparison of the AS design with adaptive designs (100,000 replicates): Results of the Monte Carlo simulation for multiplicative genetic models and the TDT with true GRR parameter $\Psi = 1.3$, a true disease allele frequency of $p = 0.2$, varying $t_{\text{interim}}$, varying assumed disease allele frequencies $p^*$, and varying $n_{\text{max}}$. Displayed are the variables for sample size ($n, n_{\text{avg}}, \text{STD}_n, \% n_2$) and the overall power. For GS design planning, $\alpha(\text{one-sided}) = 0.025$ and $\beta = 0.2$ were chosen; *italics* indicate perfect matches of planning and truth.

<table>
<thead>
<tr>
<th>$n_{\text{max}}$</th>
<th>initial plan</th>
<th>AS design</th>
<th>Changing</th>
<th>Fisher’s combination rule</th>
<th>“self-designing”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$t_{\text{interim}}$</td>
<td>$p^*$</td>
<td>$n_{\text{avg}}$</td>
<td>$\text{STD}_n$</td>
<td>power</td>
</tr>
<tr>
<td>1.5n</td>
<td>0.1</td>
<td>0.1</td>
<td>727</td>
<td>5.29</td>
<td>0.732</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>499</td>
<td>3.68</td>
<td>0.584</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>410</td>
<td>2.36</td>
<td>0.508</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>0.1</td>
<td>762</td>
<td>4.81</td>
<td>0.882</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>551</td>
<td>2.86</td>
<td>0.717</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>462</td>
<td>2.15</td>
<td>0.620</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>0.1</td>
<td>824</td>
<td>3.59</td>
<td>0.950</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>583</td>
<td>2.51</td>
<td>0.797</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>488</td>
<td>1.97</td>
<td>0.701</td>
</tr>
<tr>
<td>0.75</td>
<td>0.1</td>
<td>0.1</td>
<td>976</td>
<td>2.22</td>
<td>0.968</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>652</td>
<td>1.94</td>
<td>0.836</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>518</td>
<td>1.62</td>
<td>0.740</td>
</tr>
<tr>
<td>2.5n</td>
<td>0.1</td>
<td>0.1</td>
<td>812</td>
<td>7.18</td>
<td>0.735</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>555</td>
<td>4.15</td>
<td>0.602</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>451</td>
<td>3.07</td>
<td>0.529</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>0.1</td>
<td>831</td>
<td>6.46</td>
<td>0.887</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>613</td>
<td>4.03</td>
<td>0.784</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>514</td>
<td>3.11</td>
<td>0.642</td>
</tr>
<tr>
<td>$n_{\text{max}}$</td>
<td>initial plan</td>
<td>$t_{\text{interim}}$</td>
<td>$p^*$</td>
<td>$n_{\text{avg}}$</td>
<td>STDn</td>
</tr>
<tr>
<td>-------</td>
<td>--------------</td>
<td>--------------------</td>
<td>------</td>
<td>--------------</td>
<td>------</td>
</tr>
<tr>
<td>2.5n</td>
<td>0.50</td>
<td>0.1</td>
<td>868</td>
<td>503</td>
<td>0.957</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>541</td>
<td>343</td>
<td>0.689</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>1,093</td>
<td>641</td>
<td>1,012</td>
</tr>
<tr>
<td>4n</td>
<td>0.10</td>
<td>0.1</td>
<td>939</td>
<td>1,039</td>
<td>0.460</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>655</td>
<td>643</td>
<td>0.297</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>553</td>
<td>514</td>
<td>0.251</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>914</td>
<td>0.890</td>
<td>0.097</td>
<td>0.782</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>712</td>
<td>617</td>
<td>0.588</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>600</td>
<td>487</td>
<td>0.501</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>924</td>
<td>0.960</td>
<td>0.038</td>
<td>0.935</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>717</td>
<td>556</td>
<td>0.797</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>619</td>
<td>465</td>
<td>0.717</td>
</tr>
<tr>
<td>0.75</td>
<td>0.1</td>
<td>1,029</td>
<td>0.978</td>
<td>0.019</td>
<td>0.971</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>736</td>
<td>471</td>
<td>0.861</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>626</td>
<td>413</td>
<td>0.782</td>
</tr>
</tbody>
</table>
clearly distinguishable. Average sample sizes were either the same or slightly larger for Fisher’s combination rule; the power, however, was up to 5% larger for Fisher’s combination rule for earlier interim analyses \( t_{\text{interim}} \in \{0.1, 0.25\} \), but was up to 6% larger for the AS design and later interim analyses \( t_{\text{interim}} \in \{0.5, 0.75\} \).

With regard to the general impact of early, mid course or late interim evaluations, overall power was on average larger for later interim evaluations; notably the power increase relative to the invested sample size was particular strong between very early interim analyses \( t_{\text{interim}} = 0.1 \) and early interim analyses \( t_{\text{interim}} = 0.25 \). Expect for the “self-designing” adaptive design by Fisher (1998) for which this power increase was the smallest with 7-8%; otherwise, it ranged between 11-32%.

Finally, the exploration of varying limits for \( n_{\text{max}} \in \{1.5n, 2.5n, 4n\} \) had little impact on the results of this simulation study. As expected, larger values of \( n_{\text{max}} \) generally resulted in larger average sample sizes and larger power estimates. For \( n_{\text{max}} = 1.5n \) in comparison to \( n_{\text{max}} = 2.5n \), a relative increase in sample size between 2-12% was observed whereas for \( n_{\text{max}} = 2.5n \) and \( n_{\text{max}} = 4n \) this relative increase varied between 2-23%. Likewise the relative increase in power ranged between 0-4% or 0-6%, respectively. Note, however, that there was one exception for \( n_{\text{max}} = 4n \), very early interim analyses \( t_{\text{interim}} = 0.1 \), and overpowered initial planning \( p^* = 0.1 \). In this scenario, relatively larger average sample sizes were associated with relatively lower power. The most likely explanation for this curious behavior is the large variability in sample size reassessments indicating a large uncertainty of the average sample size estimate.

### 4.3.3 Modifying the rule for interim sample size reassessments

Since all previous simulations were based on the somewhat arbitrary formal truncation rule (4.3) for the sample size \( n_2^* \), a further simulation study was set up to explore the potential usefulness of the confidence intervals derived in chapter 3 for sample size reassessments. While the new sample size \( n_2^* \) was still limited to an upper \( n_{\text{max}} = 4n \) as in the first simulation, the option to reassess the sample size depended on whether or not all interim interval estimators, i.e., their 50%, 70%, 90%, 95% CIs, covered the true value.
for which the study was initially planned. If one of the interval estimators did not cover the true value, the sample size was reassessed as before using point estimates, conditional rejection errors, and the conditional power. In this third simulation study, again the AS design and, due to the results obtained in the last section, Fisher’s combination rule (Fisher, 1932) were explored for $t_{\text{interim}} \in \{0.1, 0.25, 0.5, 0.75\}$, i.e., varying interim evaluation times, and, as before, varying initial planning (see Table 4.3).

As before, the average sample size for the AS design was always the same or smaller than the corresponding estimate for Fisher’s combination rule while the power comparison again depended on $t_{\text{interim}}$. Fisher’s combination rule had larger power for earlier interim analyses and the AS design was superior at/after mid course of the study.

Evaluating the impact of CI choice showed that as long as 25% of the initial planning for a under- or overpowered study, are present, power and sample size variables do not vary with the coverage probability for the CIs in this simulation study. For all scenarios where planning and true genetic model matched, i.e., identical GRRs and $p = p^* = 0.2$, the impact of CI coverage probability choice differed depending on $t_{\text{interim}}$. For very early interim evaluations ($t_{\text{interim}} = 0.1$), smaller coverage values of the CIs were related to smaller average sample sizes, but also to smaller power. Running the interim analysis on 25% of the initially planned $n$ ($t_{\text{interim}} = 0.25$), larger CI coverage probabilities were associated with smaller average sample size and slightly increased power. For average sample sizes the same was true for later interim evaluations, though, then the power was slightly smaller for larger CI coverage. Note that this latter case is complementary to the settings with very early interim evaluation.

Interestingly, however, the only situation where the choice of CI coverage probability had some impact on cases with inadequate planning was the situation of the very early interim analysis for an underpowered study with an over-estimated allele frequency of $p^* = 0.3$. In this case, larger coverage probabilities were related to smaller sample sizes and the same or smaller power estimates. Thus, this case was similar to the results for late interim analyses and adequate study planning.

Note that this was just a pilot investigation - further considerations and simulation
Table 4.3: Using interval estimates for sample size reassessments (100,000 replicates): Results of the Monte Carlo simulation for multiplicative genetic models and the TDT with true GRR parameter $\Psi = 1.3$, a true disease allele frequency of $p = 0.2$, varying $t_{\text{interim}}$, varying CIs (details see text), varying assumed disease allele frequencies $p^*$, and $n_{\text{max}} = 4n$. Displayed are the variables for sample size ($n$, $n_{\text{average}}$, STD_n, $% n^2$) and the overall power. For GS design planning, $\alpha$(one-sided) = 0.025 and $\beta$ = 0.2 were chosen; italics indicate perfect matches of planning and truth.

<table>
<thead>
<tr>
<th>initial plan</th>
<th>AS design</th>
<th>Fisher’s combination rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{\text{interim}}$</td>
<td>$p^*$</td>
<td>$n_{\text{avg}}$</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95</td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95</td>
</tr>
<tr>
<td>0.75</td>
<td>0.1</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95</td>
</tr>
<tr>
<td>0.25</td>
<td>0.1</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95</td>
</tr>
<tr>
<td>0.75</td>
<td>0.1</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95</td>
</tr>
<tr>
<td>0.50</td>
<td>0.1</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95</td>
</tr>
<tr>
<td>0.75</td>
<td>0.1</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95</td>
</tr>
</tbody>
</table>
studies are needed to explore if these observations can be generalized to other deviations between planning and true genetic effect sizes and allele frequencies. Such evaluations, however, are beyond the scope of this work.

### 4.4 Real data application

Again, a real data example is presented to illustrate the new method’s potential benefit in terms of sample size reductions after the interim analysis. Whereas this chapter has focussed on family-based designs so far, the following application was done for the case-control setting. One locus and one candidate gene were examined for association with the same dichotomous phenotype. The polymorphism considered G-A (valine to isoleucine) substitution at codon 103 (rs2229616) of the *melanocortin-4 receptor (MC4R)* gene which was found to be associated with reduced risk of obesity (e.g., Geller et al., 2004). The following considerations are based on data published in Young et al. (2007). Assume one started a confirmatory candidate gene case-control study with 1:1 matching for cases defined by a body-mass-index (BMI) $\geq 30$ kg/m$^2$ and controls with a BMI $< 30$ kg/m$^2$, respectively. Furthermore, assume that the initial estimate of the minor allele frequency (MAF) for the candidate marker was $p^* = 0.04$, and the odds ratio for a dominant mode of inheritance was expected to be 0.8. Suppose that the initial planning was done for a two-tailed Pocock design (Pocock, 1977) with two equally sized stages, a significance level $\alpha = 0.05$, and an overall power of $1 - \beta = 0.80$ for the CA trend test with weights $w = (0, 1, 1)$ (see chapter 2). A sample size calculation yields a maximum of $n_{GS} = 2 \cdot 5,328 = 10,656$ individuals that need to be genotyped. Table 4.4 displays the interim and the final results.

The critical boundaries at the first stage are $c_1 = \pm \sqrt{0.5} \cdot 2.178$. The interim CA test is $T_{CA}(0.5) = -1.29$ (see chapter 2) and the interim estimates of the odds ratio for a dominant model is 0.755 (95%CI 0.551...1.032). Furthermore, the allele frequency estimator in the control group was $\hat{p} = 0.038$ (95%CI 0.031...0.046). $H_0$ cannot be rejected and without an application of the adaptive procedure the sampling of the second stage of the GS would follow. By implementing the AS procedure, one
Table 4.4: Observed numbers of genotypes for the candidate gene up to and after the interim analysis. The interim analysis was based on $n_1 = 5,328$ cases and controls which corresponds to 50% of the initial maximum GS sample size of $n_{GS} = 10,656$ individuals. The number of cases and controls that have to be sampled after the sample size adaptation is denoted by $n_2^*$. 

<table>
<thead>
<tr>
<th>marker genotype</th>
<th>up to the interim analysis $n_1 = 5,328$</th>
<th>after the interim analysis $n_2^* = 3,500$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cases</td>
<td>controls</td>
</tr>
<tr>
<td>GG</td>
<td>2,587</td>
<td>2,563</td>
</tr>
<tr>
<td>GA or AA</td>
<td>77</td>
<td>101</td>
</tr>
</tbody>
</table>

Determines $CRP_{lower} = 0.106$ and $CRP_{upper} = 4.550 \cdot 10^{-7}$ from Formula 4.2. Due to the very small value of $CRP_{upper}$, it is self-evident to abandon the corresponding side of the alternative hypothesis. With $CRP = 0.106$ and the conditional power $1 - \beta^* = 0.79$ (see Scherag et al., 2003), one may either use Formula 4.3 or software for sample size calculations (e.g., Gauderman, 2002) to determine the number of individuals that need to be genotyped after the interim analysis. Using the point estimates for adaptation, this will yield $n_2^* = 3,500$ cases and controls. The one-sided asymptotic (exact) p-value obtained for this sample was 0.065 (0.078) which in either case is smaller than the new significance level $CRP = 0.106$ indicating that $H_0$ can be rejected. In sum, the total sample size genotyped for the AS procedure was $n_{AS} = 5,328 + 3,500 = 8,828$ cases and controls. This total sample size is smaller than the maximum GS sample size $n_{GS} = 10,656$. It is even smaller than $n_{FS} = 9,593$, the corresponding sample size for a FS design.

### 4.5 Discussion

In this chapter, a method for design adaptations (Müller and Schäfer, 2004) was, for the first time, applied to the sequential recruitment and genotyping of case parent trios in candidate gene studies. Monte Carlo simulations served to explore some of its properties,
and its applicability as well as its wider scope were demonstrated by a numerical example for a case-control association study.

In most genetic studies, little is known about the genetic effects being studies beforehand (Lohmueller et al., 2003; Ioannidis et al., 2003; Hirschhorn et al., 2002; Ioannidis et al., 2001). Therefore, the main interest in introducing adaptive methods is to improve the power characteristics of the study over a wider range of true genetic effects which differ from the investigator’s pre-study estimations by adapting the sample size to interim effect size estimates.

The first Monte Carlo simulation study showed that when the true genetic effect is larger than the pre-study estimation both the GS design and the AS design will lead to considerable savings in terms of average sample sizes when compared to the FS design. If the true effects are much larger than the pre-study estimations, average sample sizes under the GS design and the AS design are nearly identical, because both designs have a high probability to stop at the interim analysis with rejection of $H_0$. When the initial estimations coincide with the true values, the GS design will still lead to considerable sample size savings, while the average sample size of AS is about 10% larger than the sample size of the FS design, paralleled by a gain in power of about 5%. When the initial estimation is larger than the true effect, the power characteristics of the AS design is superior to the FS and GS design, with a power gain of up to 10%, of course at the price of an increased average sample size. Thus, the AS design may help to compensate moderate over-estimations of the effect size or the allele frequency in the planning phase by interim sample size extensions. However, when the true genetic effects are much smaller than the investigator’s pre-study estimates, a design extension based on interim effects size estimates may frequently be impractical due to very large sample size requirements. To protect against such extreme sample sizes all three simulations studies relied on a somewhat arbitrary formal truncation rule for sample size recalculations (see Formula 4.3). Moreover, for the investigated simulation settings altering the truncation value $n_{max}$ or modifying the rule itself by using the interval estimators derived in chapter 3 had no impact on the general conclusions. If these observations hold true in general is
beyond the scope of this work. The results of others (e.g., Brannath and Bauer, 2004) at least indicate that better choices of sample size reassessment rules exist. Some proposals, however, (e.g., Tsiatis and Mehta, 2003) will imply a loss of flexibility for the most part. In contrast, the underlying statistical principle for design adaptations used for the AS design can be applied without any pre-specified sample size rule, and at any time during the course of the study, the investigator may continuously monitor the study and choose her/his own balance between the additional sample size needed and the gain in power to be achieved. Admittedly, the method’s flexibility does not excuse the investigator from doing a conscientious initial planning as was emphasized by Jennison and Turnbull (2003).

On the other hand, Posch et al. (2003) demonstrate that adaptive designs and design adaptations can also be useful when sample size reductions at the interim analysis are an option. This finding corresponds with the results of the simulation study for the cases in which the true effects were underestimated, and is also evident from the numerical example. Indeed, even the choice of the applied adaptive design or method for design adaptations such as the AS design, has an impact on the power and average sample size of the study. In the second simulation study poor power estimates were obtained for the adaptive designs by Chang (2007) and Fisher (1998) when compared to the AS design or Fisher’s combination rule. In fact, the summation of p-values for rejection of \( H_0 \) as proposed by Chang (2007) will often result in large p-values for early data evaluations and these values will already be too large for later rejections of \( H_0 \). For Fisher’s (1998) “self-designing” adaptive design it is the necessity to sample at least one more stage after adaptation, which results in an unnecessary increase in sample sizes. Moreover, if more stable interim estimators are available, e.g. at about mid-course of the initial plan, it was shown that the AS design will even outperform Fisher’s combination rule; otherwise, the multiplication of p-values will be advantageous as large initial p-values can be balanced by small p-values of the \( n_2 \) data set.

In any case, to implement the latter designs, the impact of the interim estimates has to be taken into account. The estimates applied for the family-based design are based
on the unconditional likelihood model, which assumes random mating and HWE in the case of the required case-parent trio structure. It is recommended to compare HWE and CPG estimates (see chapter 3) if possible. The precision of the interim estimates is another critical issue. Under simulation scenarios with small to moderate pre-study deviations from the true effects, $t_{\text{interim}} = 0.5$ seemed to be a reasonable choice for an interim analysis using the AS design. Moreover, it seems advisable to implement some truncation for the reassessed sample size in order to protect against instabilities of the estimates. Altering the rule itself can be considered an interesting field for future research as warranted by the interesting results obtained by using interval estimators for sample size reassessments.

In sum, the proposed adaptive procedure helps the researcher to overcome the common problem that necessary parameter values for sample size calculations are unknown beforehand. By adjusting the sample size based on interim effect size estimates derived at about half course of the initial planning, applying the AS procedure may result in more studies with an overall power that is closer to the desired power.
5 Flexible two-stage designs for genomewide association studies

5.1 Introduction

The availability of ultra-high-volume genotyping platforms (100,000 - 1,000,000 or more genotypes per sample) at a manageable cost raises the interest in conducting genomewide association studies (GWAS) for unravelling the genetics of complex traits. Numerous GWAS for a wide range of traits are being planned or conducted or they have already been published (e.g., Wellcome Trust Case Control Consortium, 2007). Despite the advances in high-volume genotyping technology, it is still prohibitively expensive to genotype hundreds of thousands of SNP markers in thousands of subjects. Therefore, sequential genotyping strategies are frequently applied and cost efficient multi-stage designs for GWAS have been propagated (Müller et al., 2007; Skol et al., 2007; Goll and Bauer, 2007; Wang et al., 2006; Wen et al., 2006; Thomas et al., 2005; Zehetmayer et al., 2005; Satagopan et al., 2004; Satagopan and Elston, 2003; König et al., 2003; Satagopan et al., 2002; König et al., 2001).

Despite the theoretical appeal of multi-stage designs of, e.g., providing strong control of the family-wise type I error rate (FWER), investigators need to comply strictly with the design specifications. For example, it is often required that markers which do not reach a nominal p-value of a certain size at interim analyses not be tested in any subsequent stage. In most cases, ignoring this leads to inflated type I error rates. Moreover, to achieve cost efficiency, genotyping of a substantial number of markers has to be limited to the first stage, e.g., about 90-95% of all SNPs must not be used for confirmatory hypothesis testing (Wang et al., 2006).
Such formal statistical rules rarely meet the practical necessities of current genetic research. Investigators may have an itch to change the number of markers or to modify the number of genotyped subjects. In practice, one may wish to base marker selection or sample size choice on other criteria than formal pre-specified statistical thresholds. This may include internal information available from interim data generated by the project itself, such as interim genetic effect size estimates (e.g., Yu et al., 2007; Scherag et al., 2003) or external information that became available in the course of a project, e.g., concerning biological and functional pathways and related candidate genes. A study by Arking et al. (2006) using ~115,000 SNPs is one example where the choice of SNP markers for further genotyping was based on both internal (smallest nominal p-values) and external information (candidate gene regions). Unfortunately, such approaches provide little information about the genomewide FWER.

In this chapter a highly flexible method for interim design modifications based on any kind of internal or external information is described which allows for controlling the genomewide FWER in a strong sense. Examples of design modifications include the alteration of the marker set used for a second study stage or the option to modify the number of individuals to be genotyped in the second stage. For this purpose, a procedure for genomewide testing of multiple genetic markers based on CRP principle (Müller and Schäfer, 2004) is developed. According to Jennison and Turnbull (2006) a particular strength of these adaptive methods lies “in coping with the unexpected, in particular responding to information that could not have been anticipated at the start of a study”. Without affecting the type I error level, the CRP principle allows for a wide range of design modifications at any time during the course of a project, and it is not restricted to a certain test statistic (see section 3.1.3).

First a notation for a case-control GWAS is described and then a more general flexible two-stage procedure for testing multiple markers is proposed. Afterwards, it is shown how to determine CRPs in practice and how to use them for design modifications with genomewide FWER control (A proof that the procedure controls the FWER in a strong sense and some more details on CRP-calculations can be found in Scherag et al. (2008)).
5 Flexible two-stage designs for genomewide association studies

Then the procedure is applied to simulated data sets in order to evaluate its statistical properties including considerations of the impact of between-marker LD. In the next step, a real data set is re-analyzed for illustration. Finally, it is shown and argued that a combination of the proposed procedure with an optimized multi-stage design will offer both gains in cost efficiency and increased flexibility.

5.2 Method

5.2.1 Notation for a case-control genomewide association study

Let $M$ denote the set of markers from which markers can be selected for genotyping and statistical testing in the study. Let $M_1 \subset M$ and $M_2 \subset M$ denote the selected set for testing in the first and second stage, respectively. For every marker $i \in M$, let $H_0^i : \theta_i = 0$ denote the null hypothesis of no association, where $\theta_i$ denotes some genetic effect size parameter. In the case of the Cochran-Armitage trend test (Sasieni, 1997) investigating a linear trend, $\theta_i$ may be the expected difference of risk allele dosage between cases and controls, where risk allele dose is 0, 1, or 2, according to the number of risk alleles in the individual (see section 2.2). Furthermore, for any marker $i \in M_1 \cup M_2$ and any sample size $k$ let $T_k^i$ denote the test statistic for testing $H_0^i$. All $T_k^i$ are assumed to be identically distributed under $H_0^i$ with cumulative distribution function $F$. The $T_k^i$ may be standardized statistics assumed to be standard normally distributed under $H_0^i$ or $T_k^i = 1 - p_k^i$ where $p_k^i$ are the marker wise $p$-values assumed to be uniformly distributed in the unit interval $[0, 1]$ under $H_0^i$. Examples are the test statistics of the standard output of the frequently used GWAS software PLINK (Purcell et al., 2007). Let $\text{Pr}_J(\cdot)$ denote the probability measure under the hypothesis that none of the markers in $J \subset M$ is associated with the trait, i.e., under the intersection hypothesis $H_0^J = \bigcap_{i \in J} H_0^i$. For a singleton set $J = \{i\}$, $\text{Pr}_i$ is used instead of $\text{Pr}_{\{i\}}$ while $E_J$ and $E_i$ denote expectations under $\text{Pr}_J$ and $\text{Pr}_i$, respectively. Throughout this chapter subset pivotality will be assumed:

$$\text{Pr}_i(T_k^i \in B) = \text{Pr}_J(T_k^i \in B)$$
for all $J \subset M_1 \cup M_2$ with $i \in J$ and for all Borel sets $B$ and for all $k$.

The researcher starts with an initial design consisting of a planned total sample size $n$, a sub-sample size $n_1 < n$ for the first stage and a marker set $M_1 \subset M$ of $m_1$ markers to be genotyped in the first $n_1$ subjects. At the planning stage, she or he furthermore specifies $m$, the maximum number of markers which can be tested, i.e. the whole set of markers $M_1 \cup M_2$ considered for testing has to consist of no more than $m$ markers. For example, for fine mapping in regions of interest the researcher may wish to add some markers in the second stage. Note that the decision on which markers to choose can be postponed until the start of the second stage. For $i \in M_1$ the planned final test of $H_0^i$ adjusted for multiple testing in $M$ is given by the indicator function $I\{T^i_n > t^{\alpha/m}\}$ where $t^\gamma := F^{-1}(1-\gamma)$ denote quantiles of $F$ and where $\alpha$ is the pre-specified level for the genomewide FWER. Of note, specification of $m$ depends on both power/sample size considerations and flexibility. Hence, usually $|M_1| < m << |M|$.

Let $X$ denote the random vector of the marker information on markers in $M_1$ collected at the first stage of the study and let $x$ denote the observed value of $X$. After the first stage, the researcher can use the whole information in $X$ to re-define the total sample size $n^*$ and to define the marker set $M_2 \subset M$, $|M_1 \cup M_2| \leq m$, of $m_2$ markers which will be genotyped in the new $n_2 = n^* - n_1$ subjects and for which a statistical test will be performed at the end of the study. Both $M_2$ and $n^*$ are random variables, $M_2 = M_2(X)$ and $n^* = n^*(X)$, but a formal rule need not be specified for the selection of $M_2$ and $n^*$. The final statistical test for all markers in $M_2$ will be based on all marker information available from both stages (Skol et al., 2006), so the following marker sets can be identified: Markers selected from $M_1$ ($M_2 \cap M_1$), new previously untyped markers ($M_2 \setminus M_1$), and abandoned markers ($M_1 \setminus M_2$). As an example imagine that $M_1$ is one of the currently available SNP arrays (Barrett and Cardon, 2006) while $M_2$ is a set of markers where genotyping is either done by using a custom array or a different technology like Matrix-assisted Laser Desorption Ionisation Time-of-Flight Mass Spectrometry.
5 Flexible two-stage designs for genomewide association studies

5.2.2 The flexible two-stage procedure

The basic idea behind the construction of the proposed flexible two-stage multiple testing procedure is to combine closed testing (Marcus et al., 1976) with the CRP approach (Müller and Schäfer, 2004). According to the closed testing principle, any single marker hypothesis \( H_i^0, i \in M_2 \), can be rejected with control of the FWER at level \( \alpha \), if all intersection hypotheses \( H_J^0 \) where \( i \in J \subset M \) can be rejected at level \( \alpha \). For any \( J \subset M \), \( J \neq \emptyset \), to test the intersection hypothesis \( H_J^0 \), according to the initial design of the study Bonferroni-type adjusted critical limits will be used, which correspond to the decision function

\[
\hat{\varphi}_J := \min\{1, \varphi_J\} \text{ where } \varphi_J := \sum_{i \in J \setminus M_1} I\{T_i^n > t_{\alpha,J}\} + \min\{m - m_1, |J \setminus M_1|\} \cdot \alpha_J, \quad \alpha_J = \frac{\alpha}{|J \setminus M_1| + \min\{m - m_1, |J \setminus M_1|\}}.
\]

After observation of \( X \) the researcher is allowed to change every \( \varphi_J \) into a new function \( \psi_J :\geq 0 \) along with the decision function

\[
\hat{\psi}_J := \min\{1, \psi_J\} \text{ such that } E_J(\psi_J|X) \leq E_J(\varphi_J|X) \text{ holds for the conditional expectations under } \Pr_J \text{ for all } J \subset M, J \neq \emptyset, \text{ according to the CRP principle. Then this flexible Bonferroni-Holm-type test procedure for the family } (H_J^0)^{J \subset M, J \neq \emptyset} \text{ maintains control of the FWER if } E_J(\varphi_J) \leq \alpha \text{ for all } J \subset M, J \neq \emptyset. \text{ The conditional expectations of decision functions are called conditional rejection probabilities.}
\]

Throughout this chapter, let \( \text{CRP}_J(c) \) denote the sum of the conditional rejection probabilities of all markers in \( J \subset M, J \neq \emptyset \), for an arbitrary critical limit \( c \) at the planning stage, i.e.,

\[
\text{CRP}_J(c) := \sum_{i \in J \cap M_1} \mathbb{E}_i \{I\{T_i^n > c\}|X = x\} + \min\{m - m_1, |J \setminus M_1|\} \cdot \left(1 - F(c)\right).
\]

In the new procedure, after the first stage has been completed, \( \varphi_J \) is replaced by

\[
\psi_J := \sum_{i \in J \cap M_2} I\{T_i^{n_\ast} > c_J\} \text{ if } J \cap M_2 \neq \emptyset \text{ and by } \psi_J := 0 \text{ if } J \cap M_2 = \emptyset, \text{ where } T_i^{n_\ast} \text{ is the test statistic for the single marker } i \in M_2 \text{ based on the modified total sample size } n_\ast, \text{ i.e. based on } n_\ast \text{ if } i \in M_1 \cap M_2 \text{ and based on the } n_2 \text{ new subjects if } i \in M_2 \setminus M_1 \text{ where } c_J \text{ is a common critical limit for markers in } J.
\]

After these design modifications let in the following denote by \( \text{CRP}_J^*(c) \) the sum of the conditional rejection probabilities of all markers in \( J \subset M, J \neq \emptyset \) at an arbitrary critical limit \( c \), i.e.,

\[
\text{CRP}_J^*(c) = \sum_{i \in J \cap M_2} \mathbb{E}_i \{I\{T_i^{n_\ast} > c\}|X = x\}.
\]
For every $J \subset M$, $J \neq \emptyset$, define $c_J := \inf \{ c \mid \text{CRP}_J^*(c) \leq \text{CRP}_J(t^{\alpha_J}) \}$. In Scherag et al. (2008) a mathematical proof is given that $E_J(\psi_J|X) \leq E_J(\varphi_J|X)$ for all $J \subset M$, $J \neq \emptyset$. In this way a flexible closed testing procedure at level $\alpha$ is defined.

In addition, it is demonstrated in Scherag et al. (2008) that $H_0^I$ can be rejected with control of FWER at level $\alpha$ for any marker $i \in M_2$ for which $T_{n^i} > \hat{c}$ holds true. If the total sample size is not modified, i.e., $n^* = n$ the common critical limit $\hat{c}$ is defined by

$$\hat{c} := \inf \{ c \mid \text{CRP}_{M_2}^*(c) \leq \min_{J \subset M, J \supset M_2} \text{CRP}_J(t^{\alpha_J}) \}.$$  \hspace{1cm} (5.1)

If the researcher modifies the total sample size, a further adjustment of the critical limit $\hat{c}$ may be necessary (see Scherag et al., 2008).

The $\min$ in condition (5.1) can be found by the following algorithm.

Initialize: $\min = \text{CRP}_{M_2}(t^{\alpha/m_2})$

DO $h = 1$ TO $|M_1 \setminus M_2| + m - |M_1 \cup M_2|$

[ Identify the marker set $J \subset M \setminus M_2$, $|J| = h$ with $\text{CRP}_{M_2 \cup J}(t^{\alpha/(m_2+h)})$

$= \min_{H \subset M \setminus M_2, |H| = h} \text{CRP}_{M_2 \cup H}(t^{\alpha/(m_2+h)})$.]

Calculate $\text{CRP}_{M_2 \cup J}(t^{\alpha/(m_2+h)})$ given the initial planning with $n$ individuals.

IF $\text{CRP}_{M_2 \cup J}(t^{\alpha/(m_2+h)}) < \min$, THEN $\min = \text{CRP}_{M_2 \cup J}(t^{\alpha/(m_2+h)})$

Note that this algorithm is of linear complexity if $\arg \min_{H \subset M_1 \setminus M_2, |H| = h} \text{CRP}_{M_2 \cup H}(c)$ does not depend on the choice of $c$, which is the case in most practical situations where the same number of individuals is genotyped in the first stage for all markers in $M_1$. In this case, before starting the DO loop the markers in $M_1 \setminus M_2$ are ordered by the values of $e_i = E_i \{ T_n^i > c \} |X = x \}$, for an arbitrary choice of $c$. In every step of the DO loop, the marker set $J$ is then generated by adding the marker with the smallest $e_i$ to the set $J$ identified in the last step. In a certain step of this loop, the $m - |M_1 \cup M_2|$ placeholders for additional markers which finally were not included in $M_2$ can also be added to the set $J$. 

76
The common critical limit c for the final test of each marker can be determined by a bi-sectional search over c in equation (5.1). Note that $E_i(I\{T_{n^*} > c\}|X = x) = 1 - F(c)$ for markers $i \in M_2 \setminus M_1$ as their $T_{n^*}$ and $X$ are stochastically independent.

If the null hypotheses for a subset of $M_2$ can be rejected at a critical limit of $\tilde{c}$, Holm-steps may follow in order to reject null hypotheses for further markers in $M_2$ until no additional null hypotheses can be rejected. For such a step, $M$ (and accordingly $M_1$ and $M_2$) is reduced by the markers for which the null hypotheses could be rejected and then the algorithm is repeated for determination of a lower critical limit than $\tilde{c}$.

**5.3 Simulation studies**

To assess the performance of the proposed method, a set of $M_1$ of 1,000,000 SNP markers with varying degrees of LD was simulated. LD was assumed to decay exponentially as the intermarker distance increases, mimicking the erosion of the initial association by recombination events (Hartl and Clark, 2007). This means that when $r^2 = 0.9$ between two adjacent markers, the value will be $\approx 0.6$ between SNPs that are 5 markers apart. For sake of simplicity it is assumed that the initial planning was done according to the suggestion of Witte et al. (2000) where the FWER is controlled by the Bonferroni method for $\alpha = 0.05$ (two-sided), desiring a minimum power of $1 - \beta = 0.8$ to detect a standardized genetic effect of $\vartheta_i = 0.25$ in units of standard deviations. Under the alternative hypothesis either 1 or 10 SNP markers were associated with the same $\vartheta = 0.25$; for the remaining SNPs $\vartheta_i = 0$. The latter null markers can therefore be used for estimating FWERs. The interim data examination was performed after genotyping 50% of the initially planned total sample size. Either 1,000 or 10,000 SNP markers of the initial set $M_1$ were selected for genotyping in the remaining 50% of the total sample size (see Table 5.1). Note that this situation matches both current practice as well as the requirement for the linear complexity of the suggested algorithm. For all the simulations, the choice of SNPs for further genotyping was based on the size of their interim test statistics. In a second simulation, the influence of adding 100 new, previously untyped SNPs to the second stage marker panel was explored assuming that
5 SNPs were truly associated in the original panel and another 5 were associated in the newly added markers (see Table 5.2). The number of individuals genotyped in the second stage was varied in a third simulation. Both a reduction of 25% and a 25% enlargement relative to the initial, total n were investigated (see Table 5.3). For each scenario, 1,000 Monte Carlo replicates were generated.

As intended, the observed FWERs, i.e., the probabilities of one or more false rejections among all SNPs with \( \theta_i = 0 \), was controlled at a level \( \alpha = 0.05 \). All FWERs in Table 5.1-5.3 range between 0.002 and 0.058 for varying levels of LD between two adjacent markers with more conservative values for increased marker dependency. In particular the FWER was controlled at the level \( \alpha = 0.05 \) for both the modification of the marker panel (see Tables 5.1-5.2) and the modification of the sample size for the second stage (see Table 5.3).

All tables also display power estimates, i.e., probabilities to identify either one (minimum power) or \( \geq j \) of the truly associated SNPs. Except for the cases of a substantial marker reduction (see Table 5.1 for \( |M_2| = 1,000 \)) or a substantial sample size reduction (see Table 5.3), the Monte Carlo power estimate of the proposed method was always the same as or larger than the power for the initial design with a power of \( 1 - \beta = 0.8 \) to detect the truly associated SNP with the smallest signal of interest. For all cases, the initial plan required genotyping of \( |M_1| \) SNPs in n subjects. For the proposed flexible method genotyping was reduced to \( |M_1| \) SNPs in \( 0.5 \cdot n \) subjects + \( |M_2| \) SNPs in \( 0.5 \cdot n \) subjects. Even when an optimal multistage plan (e.g., Müller et al., 2007; Wang et al., 2006) is not applied, the costs of such an equally powered project can be approximately halved for, e.g., \( |M_1| = 1,000,000 \) and \( |M_2| = 10,000 \) assuming constant per genotype and per phenotype costs in both stages.

Finally, in a fourth scenario, the sample sizes of the simulation study was planned with an cost optimal multistage design (Wang et al., 2006). The simulation was such that either the cost optimal design was continued after the interim examination or the flexible design was implemented with the same marker selection criterion as in the case of the optimal design. For the flexible design, however, the initial critical values were
Table 5.1: Monte Carlo FWER and power based on 1,000 replicates for the new method with an initial set of $|M_1| = 10^6$ SNP markers, varying degrees of LD between markers, and varying set sizes $|M_2|$ for the second stage.

| $|M_2|$ | $r^2$ | 1 SNP associated | 10 SNPs associated |
|--------|-------|------------------|--------------------|
|        |       | FWER power       | FWER power to detect $\geq j$ of 10 associated SNPs |
|        |       |                  | $j = 1$ $j = 2$ $j = 3$ $j = 4$ $j = 5$ $j = 6$ $j = 7$ $j = 8$ $j = 9$ $j = 10$ |
| 1,000  | 0.00  | 0.047 1.000 0.986 0.938 0.794 0.556 0.262 0.060 0.000 0.000 0.000 |
|        | 0.50  | 0.028 1.000 0.997 0.986 0.940 0.804 0.550 0.281 0.071 0.000 0.000 |
|        | 0.90  | 0.023 1.000 0.996 0.980 0.936 0.812 0.588 0.300 0.074 0.000 0.000 |
|        | 0.99  | 0.004 1.000 0.999 0.996 0.985 0.923 0.819 0.592 0.275 0.060 0.000 |
| 10,000 | 0.00  | 0.056 1.000 0.997 0.992 0.957 0.862 0.655 0.362 0.103 0.000 0.000 |
|        | 0.50  | 0.055 1.000 0.995 0.958 0.876 0.674 0.375 0.112 0.000 0.000 0.000 |
|        | 0.90  | 0.028 1.000 0.999 0.994 0.961 0.885 0.691 0.370 0.089 0.000 0.000 |
|        | 0.99  | 0.003 1.000 0.999 0.977 0.878 0.665 0.387 0.105 0.000 0.000 0.000 |
Table 5.2: Inclusion of previously untyped markers: Monte Carlo FWER and power based on 1,000 replicates for the new method with an initial set of $|M_1| = 10^6$ SNP markers, varying degrees of LD between markers, varying set sizes $|M_2|$ and the inclusion of 100 previously untyped markers for the second stage.

| $|M_2|$ | $r^2$ | 5 SNPs associated in $M_1$ and 5 SNPs associated among the previously untyped SNPs |
|-------|-------|----------------------------------------------------------------------------------|
|       |       | FWER | power to detect $\geq j$ of 10 associated SNPs | j = 1 | j = 2 | j = 3 | j = 4 | j = 5 | j = 6 | j = 7 | j = 8 | j = 9 | j = 10 |
| 1,000 | 0.00  | 0.50 | 0.00  | 0.050 | 0.00  | 0.998 | 0.971 | 0.840 | 0.558 | 0.224 | 0.060 | 0.007 | 0.001 | 0.000 |
|       | 0.50  | 0.029| 0.999 | 0.993 | 0.946 | 0.809 | 0.526 | 0.215 | 0.063 | 0.016 | 0.003 | 0.000 |
| 100   | 0.90  | 0.028| 0.999 | 0.991 | 0.948 | 0.799 | 0.491 | 0.211 | 0.085 | 0.024 | 0.005 | 0.001 |
|       | 0.99  | 0.006| 1.000 | 0.993 | 0.936 | 0.759 | 0.425 | 0.194 | 0.150 | 0.093 | 0.050 | 0.013 |
| 10,000| 0.00  | 0.50 | 0.00  | 0.050 | 1.000 | 0.997 | 0.981 | 0.853 | 0.603 | 0.272 | 0.085 | 0.009 | 0.000 | 0.000 |
|       | 0.50  | 0.058| 0.999 | 0.995 | 0.978 | 0.880 | 0.626 | 0.297 | 0.086 | 0.012 | 0.005 | 0.001 |
| 100   | 0.90  | 0.028| 0.999 | 0.993 | 0.958 | 0.845 | 0.578 | 0.265 | 0.121 | 0.045 | 0.013 | 0.002 |
|       | 0.99  | 0.002| 0.999 | 0.994 | 0.950 | 0.791 | 0.488 | 0.204 | 0.152 | 0.108 | 0.064 | 0.024 |
Table 5.3: Sample size modification: Monte Carlo FWER and power (1 SNP associated) based on 1,000 replicates for the new method with an initial set of $|M_1| = 10^6$ SNP markers, varying degrees of LD between markers, varying set sizes $|M_2|$ as well as a modified sample size for the second stage (for details see text).

<table>
<thead>
<tr>
<th>$M_2$</th>
<th>$r^2$</th>
<th>sample size reduced by 25%</th>
<th>sample size increased by 25%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FWER</td>
<td>power</td>
</tr>
<tr>
<td>1,000</td>
<td>0.00</td>
<td>0.049</td>
<td>0.529</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.045</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.020</td>
<td>0.482</td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>0.006</td>
<td>0.493</td>
</tr>
<tr>
<td>10,000</td>
<td>0.00</td>
<td>0.046</td>
<td>0.528</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.057</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.028</td>
<td>0.496</td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>0.002</td>
<td>0.495</td>
</tr>
</tbody>
</table>

Table 5.4: Comparison with optimal two-stage designs: Monte Carlo FWER and power based on 10,000 replicates. The parameters for the optimal case-control design with constant per genotype costs are derived from Wang et al. (2006, Table III.) aiming at a power of 90% for one truly associated SNP with a MAF $p = 0.20$ and a (multiplicative) genotype relative risk of $\psi = 1.5$ and $\phi = 0.5$. The overall type I error rate is $0.05/|M_1|$ (two-sided) and $n$ is the total number of individuals genotyped.

| $|M_1| (x10^6)$ | optimal design by Wang et al. (2006) | flexible design |
|--------------|-------------------------------------|-----------------|
|              | $\alpha_1$ | $\alpha_2 (x10^{-5})$ | $\pi$ | $n$ | FWER | power | FWER | power |
| 30           | 0.070       | 0.2400             | 0.209 | 3,292 | 0.0501 | 0.8859 | 0.0480 | 0.8855 |
| 100          | 0.064       | 0.0720             | 0.200 | 3,536 | 0.0530 | 0.8885 | 0.0505 | 0.8876 |
| 500          | 0.056       | 0.0150             | 0.187 | 3,894 | 0.0499 | 0.8897 | 0.0449 | 0.8893 |
| 1,000        | 0.053       | 0.0072             | 0.187 | 3,968 | 0.0537 | 0.8919 | 0.0509 | 0.8915 |
Bonferroni-corrected and re-determined at the interim examination using the algorithm described in section 5.2.2. Table 5.4 shows the results of this comparison. For both designs the FWER was controlled at the pre-specified level. More importantly, however, not modifying the initial optimal plan and continuing with the flexible procedure had no impact on the intended power of the study, thus showing that there is no price to be paid for the option of design modifications and that the implementation of the algorithm for the flexible method works reasonably well.

5.4 Real data application

To demonstrate the applicability of the proposed method, the data of a GWAS for adolescent and childhood obesity (Hinney et al., 2007) comprising 487 extremely obese cases and 442 healthy underweight controls was re-analyzed. For the purpose of this chapter it is assumed that the study was initially planned according to the cost efficient two-stage designs proposed by Wang et al. (2006). Accordingly, all $|M_1| = 440,794$ SNPs have to be evaluated in 16.1% of all case-control pairs and only SNPs with p-values smaller than 0.055 (two-sided) should be genotyped in the remaining individuals. Finally, in order to reject $H_0$ the markerwise combined p-value of both stages has to be smaller than $1.6 \times 10^{-7}$ (two-sided). Whereas “genomewide significance” was observable for SNPs in the Fat Mass and Obesity Associated Gene \textit{(FTO)} in the original publication of the complete data using a fixed sample size (Hinney et al., 2007), none of the SNPs would have been significant using the two-stage design. Thus, the investigators would have missed this well replicated genetic association (e.g., Scuteri et al., 2007) due to insufficient power at the interim analysis. By contrast, the new flexible method would at least have provided an option to incorporate the candidate gene information for the \textit{FTO} gene which became available at the time of the GWAS genotyping. Including \textit{FTO} SNPs in the marker panel for the second stage, applying the new flexible method might have even resulted in a genomewide significant result for the markers. Table 5.5 is a summary of the data for the SNP rs1121980 in \textit{FTO} which resulted in a combined p-value of $5.95 \times 10^{-8}$ (one-sided) which is smaller than the newly determined markerwise
Table 5.5: Observed numbers of genotypes for the marker rs1121980 in the FTO gene up to and after the interim analysis; for three individuals genotypes were missing.

<table>
<thead>
<tr>
<th>marker genotype</th>
<th>up to the interim analysis ( n_1 = 149 )</th>
<th>after the interim analysis ( n_2 = 777 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cases</td>
<td>controls</td>
</tr>
<tr>
<td>CC</td>
<td>17</td>
<td>26</td>
</tr>
<tr>
<td>CT</td>
<td>42</td>
<td>30</td>
</tr>
<tr>
<td>TT</td>
<td>19</td>
<td>15</td>
</tr>
</tbody>
</table>

significance level of \( 1.51 \times 10^{-7} \) (two-sided). Details on these calculations are provided in Scherag et al. (2008).

### 5.5 Discussion

This chapter deals with a solution to a dilemma that is evident in all properly planned genomewide association projects: Whenever changes need to be made during the course of the project, investigators will have to stick to their original design. Otherwise, control of the genomewide FWER will not be possible.

Some have argued that control of the FWER for GWAS is not necessary as they are mainly useful for generating hypotheses (e.g., Boehringer et al., 2000). This is in contrast to current reporting practice. Indeed, published reports of GWAS usually display tables of the smallest nominal p-values (e.g., Wellcome Trust Case Control Consortium, 2007). These p-values may, however, be heavily biased due to intensive data mining steps, e.g., exploring multiple available phenotypes or subgroups. Here it is argued in favour of controlling the FWER in a strong sense even in initial GWAS, instead of reporting uncorrected p-values. This will provide investigators with more confidence in their results which is a necessary requirement to justify subsequent molecular work within some interesting regions and genes.

All proposed multi-stage designs (e.g., Skol et al., 2007; Goll and Bauer, 2007; Müller
et al., 2007; Wang et al., 2006) also provide some measure of type I error control. However, this error control comes at the price of pre-defined and rigid study designs. Usually both the numbers of markers and individuals allocated to the respective stages will be fixed and cannot be changed during the course of the study. Moreover, investigators will have to accept formal statistical boundaries for marker selection in the second stage.

In this chapter a new, more flexible method for genomewide association (multi-stage) studies is introduced. This method relies on the framework of closed testing (Marcus et al., 1976) and uses the CRP principle for design adaptations (Müller and Schäfer, 2004). In particular, it was proposed to determine a (minimum) sum of markerwise conditional type I error rates at any unplanned interim data examination. Given this sum, numerous design modifications can be realized and control of the FWER in a joint analysis will still be possible. Furthermore, the proposed algorithm allows for testing the respective null hypothesis of each individual marker in the final analysis.

As in almost all genetic studies little is known about the genetic effects under study beforehand (Ioannidis et al., 2006), the suggested procedure will be of great value for situations in which new information becomes available during the course of a project. For the candidate gene association studies, the results of chapter 4 and Scherag et al. (2003) have shown that the power characteristics of a study can be improved when the true genetic effects differ from the investigator’s pre-study guesses, e.g., by adapting the sample size to interim effect size estimates (Yu et al., 2007). Such flexibility, however, does not excuse the investigator from conscientious initial planning as pointed out by Tsiatis and Mehta (2003). This is particularly true for the situation of adding new markers at the interim evaluation. The set size choice must be data independent and a too large set will result in a loss of power. Moreover, given an ideal world of investigators being able to correctly guess the true allele frequencies, genetic effect sizes and the mode of inheritances, an optimized cost efficient multi-stage design would always be the most efficient statistical choice. The described re-analysis of a GWAS (section 5.4) demonstrates that this may not always be the case in real life.

Finally, one should note the computational requirements for performing the calcula-
tions for the proposed algorithm. A basic requirement is the availability of fast and precise numerical integration functions. In a practical situation, calculating chromosome-wise statistics may be preferable as this will require the integration for less extreme tails. In the simulation study, one replicate of 1,000,000 SNPs was calculated in less than a CPU minute without requiring a high-performance computer (AMD Athlon\textsuperscript{TM} 64 3,500+, 2,200 Mhz, 1 GB RAM).

In sum, the new method provides increased flexibility for GWAS. Combined with one of the recently proposed cost efficient multi-stage designs, genomewide FWER control can be provided, costs can be reduced and data can be analysed jointly (Skol et al., 2006). Besides allowing for more flexible marker selection, the method addresses the previously unanswered problem of how to allow for genotyping of new markers not genotyped in the first stage, e.g., SNPs that are not available on fixed SNP arrays, while still providing FWER control in a strong sense. This is an option to deal with the criticism of indirect LD mapping (Terwilliger and Hiekkalina, 2006). Moreover, the method will be useful for sample size increases to improve their detectability of rare variants (Yu et al., 2007). Another application of the new procedure, and perhaps the most challenging, is the option to sequentially combine different test statistics investigating the same underlying parameter for example in a case-control and a family-based data setting (Weinberg and Umbach, 2005). Finally, one should note that GWAS are just one possible application of the proposed procedure. In microarray experiments (Dupuy and Simon, 2007) involving the expression levels of thousands of genes or whenever multiple hypotheses testing in high dimensional data sets is an issue, the new method may be useful.
6 General discussion

The purpose of this thesis was an introduction of design adaptation methods in the field of statistical genetics. Simulations and real data applications were done for both a candidate gene and a genomewide association setting. Moreover, case-control as well as family-based designs were explored. In particular, this work comprised three topics - (1) design adaptations for candidate gene association studies (chapter 4) employing interim effect size estimators (2) that were introduced in chapter 3. In chapter 5, a new flexible two-stage design for genomewide association studies was presented and evaluated (3).

Whereas each chapter already ended with a discussion, this general discussion will outline prospects for future research. Contrary to the structure of this work, it starts with methods for design adaptation in genomewide association studies due to the dominance of this approach in human genetics of complex diseases or traits. Extracting information from the data sets and producing valid, reproducible findings is one of the challenges of bioinformatics, statistical genetics, and genetic epidemiology. Consequently, this general discussion will comment on these issues including a more general view which addresses problems beyond the actual scope of this work. Next, implications and extensions for design adaptations and candidate gene studies are provided. The work concludes with considerations for estimators of genetic effect sizes and their relevance for gene characterization in the context of design adaptations. This modified order was chosen due to the fact that most future candidate gene investigations will arise from genomewide association scans. Note the similarity to genomewide linkage scans and the subsequent fine-mapping by association studies. Similar to the general scope of this work, this general discussion will focus on the challenges for linkage disequilibrium mapping in common complex traits.
6.1 Design adaptation methods for genomewide association studies

Technological progress allows for conduction of genomewide association studies of common SNPs in order to identify DNA sequence variants related to disease risk. Despite substantial differences in technology, all available first-generation SNP panels rely on indirect LD mapping and offer similar levels of genomic coverage for common variants (Barrett and Cardon, 2006; Jørgenson and Witte, 2006b). For this reason, planning a genomewide association study is less constrained by LD coverage considerations and more strongly requires knowledge about the genetic model for the investigated complex trait. In the majority of cases, however, such information is missing (e.g., Ioannidis et al., 2006). A procedure for conducting such a study despite these uncertainties was developed in chapter 5. Using arbitrary information for ranking genetic markers, it was shown how to adapt the subsequent part of a genomewide study while controlling the genomewide type I error rate (FWER in a strong sense). Compared to all previously proposed genomewide multi-stage designs where formal statistical rules have to be met, the procedure addresses the practical need for increased flexibility of ongoing genetic research using genomewide SNP panels (e.g., Hampe et al., 2007; Sladek et al., 2007; Frayling et al., 2007; McPherson et al., 2007). Moreover, the requirement to comply with formal statistical rules may be one of the main reasons why truly sequential study plans with a joint final analysis (Skol et al., 2006) are rarely found in practice.

Increased flexibility for genomewide association studies was introduced using a general principle for design adaptations (Müller and Schäfer, 2004). Compared to adaptive designs (see Victor and Hommel (2007) or chapter 3), the application of this method is in some respect advantageous as it results in no loss of efficiency when compared to the optimal design if no design alterations are conducted. Thereby, the proposed procedure also indirectly addresses one of the criticisms of flexible designs in the field of clinical trials (Burman and Sonesson, 2006). With regard to the prospects of the CRP principle in genetics, extensions to include gene-gene (e.g., Evans et al., 2006; Marchini et al., 2005) or gene-environment interactions (e.g., Hunter, 2005) as well as tests of
multilocus genotypes or haplotypes (e.g., Schaid, 2006; Clayton et al., 2004; Ott and Hoh, 2001) are conceivable. Within the multiple testing framework presented in chapter 5, this will increase the number of tests considered and requires to specify the respective test statistics at the respective stage of the design. Furthermore, conditional rejection probabilities have to be computable which is a fundamental requirement for all tests considered. In any case, to obtain a reproducible genetic association, it is the a-priori probability of association that needs to be accounted for rather than the number of tests. As a simple consequence of Bayes’ Theorem, even for the situation of a small false positive probability, most positive results will turn out to be false (e.g., Moonesinghe et al., 2007; Zaykin and Zhivotovsky, 2005; Ioannidis, 2005). Adopting for instance a pre-study probability of 0.1% for a true association and conducting a study with 90% power and a significance level $\alpha = 0.05$ leads to a posterior probability of 0.02 that a significant test result is a true positive finding as shown by Ott (2004). No statistical procedure or modified statistical significance criterion like the “false discovery rate” (Benjamini et al., 2001) or the “q-value” (Storey and Tibshirani, 2003) can circumvent this problem. In fact, both proposals assume or work well if the amount of true alternative hypotheses compared to all hypothesis tested is relatively large.

More importantly, systematic errors, including insufficient statistical planning or simplified assumptions and beliefs, will be a greater danger to the validity of the results generated from genomewide association scans (e.g., Terwilliger and Hiekkalinna, 2006; Ioannidis, 2005; Spence et al., 2003). In this regard, initiatives like the “Human Genome Epidemiology Network” (Ioannidis et al., 2006, 2005) require support for addressing some of these problems.

Despite the caveats outlined, some genomewide association projects will and have resulted in reproducible associations between allelic variants and complex disease (e.g., Frayling et al., 2007; Young et al., 2007; Lyon et al., 2007; Helgason et al., 2007). For some complex traits, however, tagging common variation by indirect mapping would be inappropriate. The phenotype may be more likely to be the product of multiple potentially interacting, rare variants as proposed for the phenotype extreme human
body weight (Ahituv et al., 2007). If detecting more of these variants is of interest, one idea is to re-analyze existing genomewide linkage scans using meta-analysis (e.g., Saunders et al., 2007; Dempfle and Loesgen, 2004) or meta-regression techniques (van Houwelingen et al., 2002). Afterwards, genotyping at higher density than the available SNP panels may be done within the most promising chromosomal regions using for instance Illuminas custom iSelect Bead Chip at a density of 19,760 SNPs or 38,000 SNPs within an arbitrary region (Steemers and Gunderson, 2007; Fan et al., 2006). Another idea is the generation of gene-centric genomewide association panels (as proposed by Jorgenson and Witte, 2006b). Finally, complete genome sequencing at manageable costs (Bentley, 2006) would become an option. Given raw data of about $3 \times 10^9$ genotypes for each individual, the challenge of how to extract meaningful information from this data would even be greater. Based on previous experiences, it can be expected that technology will be the driving force for the future developments of statistical methods. Such procedures will have to address the problem of how to analyze complete genome sequences of many individuals. Again, flexible multi-stage designs may be a useful tool to control the type I error rate. For this scenario, one may imagine a situation where some bioinformatic data mining tool to model genetic pathways, to detect DNA sequence motifs or to jointly model high-density SNP as well as gene expression data (Wang et al., 2007; English and Butte, 2007; Dixon et al., 2007; D’haeseleer, 2006; Schadt et al., 2005) is used for marker selection at an interim evaluation. Furthermore, Bayesian considerations (e.g., Wakefield, 2007; Wacholder et al., 2004) might be used to rank the first stage markers. The design modification method proposed in chapter 5, however, may be a tool for type I error control of the complete design.

### 6.2 Design adaptation methods for candidate gene investigations

After the initial genomewide investigations, follow-up work will focus on some of the interesting, presumably significant, genetic associations. Prior to functional characterization of the genetic variants, replications of these initial results are urgently needed.
6 General discussion

These efforts will usually require access to samples larger than the one used for the initial report (Ioannidis et al., 2003). While for genomewide scans control of the type I error rate may be of minor importance, it is a crucial requirement for these studies. Again, multi-stage designs with an option for design adaptation may be a cost-efficient and flexible alternative compared to the current practice of demonstrating that “presence” or “absence” of a genetic association in a broad variety of genotyped samples often collected under very different ascertainment schemes (e.g., Lyon et al., 2007).

Whether or not adjusting the sample size on the basis of observed genetic effect size estimates at interim analysis is a good proposal (see chapter 4), does depend on study design parameters such as “time of interim analysis” and the “choice of stopping criteria”. Some of these features were evaluated in chapter 4. For the field of clinical trials, similar questions were recently addressed by Jahn-Eimermacher and Hommel (2007) - within the field of genetics, investigations need to be extended. Unlike in clinical trials where one, two or three hypotheses are tested, confirmatory genetic candidate gene studies will still include a larger number of hypotheses. The literature dealing with treatment selection for multi-arm Phase II clinical trials within the context of adaptive designs (e.g., Bretz et al., 2006) may, thus, be a source of future research. Furthermore, for this more limited number of hypotheses one may combine tools for numerical integration of multivariate normal distributions (e.g., Conneely and Boehnke, 2007) with design adaptation techniques.

While in chapter 4 interim estimates of genetic effects were used for sample size modifications, another prospect of design adaptations may deal with switching the test statistic at the interim evaluation. As an example, assume that a “model-free” chi-square based test statistic was foreseen for investigating the 2 x 3 contingency table of a confirmatory case-control design. The interim data, however, were in good agreement with a dominant mode of inheritance. This may result in a decision to use the CA trend test with weights reflecting this new information. Furthermore, this change in the test statistic may also be useful for modifying the sample size of the remaining study. Such adaptations may even lead to more stable sample size estimates than the ones obtained in chapter 4, but
such investigations are beyond the scope of the current thesis.

### 6.3 Limits and prospects of gene characterization using estimates of genetic effects

In chapter 4, interim effect size estimators were used for sample size reassessment in a candidate gene association study. Furthermore, the usefulness of interval estimators, derived in chapter 3 (for details see Scherag et al., 2002), to indicate the necessity of sample size reassessments was explored. Initial simulations at least warranty a need for further investigation.

With regard to the role of interim estimators, it was argued that the estimation of genetic effect sizes provides more information than the mere result of a statistical test such as the TDT (Spielman et al., 1993). Yet, one should also note that using the estimates proposed in chapter 3 implies that marker and trait locus coincide. Otherwise, LD between the marker SNP and the disease locus will lead to effect size estimates biased towards the null hypothesis (Franke et al., 2005; Zondervan and Cardon, 2004). If the same data set was used to identify the variant of interest (Garner, 2007; Lohmueller et al., 2003; Göring et al., 2001), an upward bias called the “winners curse” is more likely to be present. This is due to at least two reasons: First, samples that are used for the initial association mapping are often collected to oversample affected individuals relative to their frequency in the population. As a consequence, genetic risk estimates may not be transferable to the general population. Second, for the situation of a genomewide association study selecting the most extreme test statistics such as the smallest p-values is often equivalent to selecting the most extreme genetic effect size estimates. Hence, the estimates must also be adjusted with respect to this multiplicity.

To circumvent this problem, estimation of genetic effect sizes may be done in independent, appropriately powered confirmatory studies (Garner, 2007). Alternatively, these problems may be addressed by developing bias-corrected estimators which is an active area of research (Yu et al., 2007; Zöllner and Pritchard, 2007; Huang and Lin, 2007). Note that unbiased estimation of genetic effects is also of crucial importance for sample
size planning of confirmatory studies. Moreover, reliable gene characterization is a first step towards clinical applications - the implementation of genetic marker information for prognostic, diagnostic or interventional purposes. A vision for such aspects of genomics research can be found in Collins et al. (2003). Zeggini and McCarthy (2007) and Janssens et al. (2006c,a,b), however, discuss the merits of predictive genetic testing for non-insulin dependent diabetes mellitus given the real data example of the transcription factor 7-like 2 (TCF7L2) haplotypes (e.g., Grant et al., 2006; Helgason et al., 2007) and within a wider scope.

More than 10 years after the proposal of Risch and Merikangas (1996) to initiate genomewide association studies, the “harvest” seems to begin as indicated by weekly reports of such studies in high-ranking journals (e.g., Wellcome Trust Case Control Consortium, 2007; Frayling et al., 2007; Frayling, 2007; McPherson et al., 2007; Hampe et al., 2007; Sladek et al., 2007; Herbert et al., 2006; Arking et al., 2006; Klein et al., 2005). Exploiting these findings in order to understand pathological processes (e.g., Bourgain et al., 2007) and to address clinical goals, however, is a far more difficult task that will vary substantially between the complex clinical phenotypes investigated. While the identification of associated markers with smaller, consistent effects will continue, cooperations with experts in clinical, cellular, animal and molecular biology will be needed in order to work out the mechanisms behind these associations and to bridge the gap between a statistical finding and its clinical implication.
7 Summary

Genetic association studies have become the most widely used gene mapping tool for the identification of disease susceptibility loci of complex common traits or diseases. In order to obtain sufficient power at a certain significance level (type I error risk), these analyses require a complete pre-specification of the total number of individuals to be sampled and genotyped. However, in most of these studies little information about the genetic effect size is available beforehand and thus it is difficult to calculate a reasonable sample size. By addressing these problems, this thesis aims at introducing, extending, and evaluating statistical methodology on design adaptations for genetic association studies.

In particular, it is shown how a confirmatory candidate gene association study can be planned and analyzed given the mentioned uncertainties. For this purpose, an adaptive group sequential procedure is developed. If no rejection of the null hypothesis is possible at the interim analysis, the design of the subsequent study part can be modified depending on interim data. As an example, sample size reassessment may be done using interim effect size estimates developed by the author of this thesis, as well. Finally, a new flexible two-stage design for genomewide association studies is presented. While providing strong control of the genomewide, family-wise type I error rate, the new design might also be more cost-efficient due to greater flexibility in comparison to all previously suggested designs. Examples of which are the possibility to base marker selection upon biological criteria instead of statistical criteria or the option to modify the sample size at any time during the course of the project.

Both the candidate and the genomewide association designs are evaluated using simulated and real data sets. Finally, prospects and limits of design adaptation methods and estimators of genetic effects in genetic association studies of complex traits are discussed.
8 Zusammenfassung


8 Zusammenfassung

References


References


References


References


References


References


References


Lebenslauf

Persönliche Daten

André Scherag
Rellinghauserstr. 266
45136 Essen

Tel.: (0201) 266 7405
E-Mail: andre_scherag@hotmail.de

geb. am 07. 12. 1974 in Bad Hersfeld
ledig, deutsch

Berufserfahrung

seit 2007 Wissenschaftliche Mitarbeiter am Institut für Medizinische Informatik, Biometrie und Epidemiologie (Prof. Dr. Jöckel), Universität Duisburg-Essen

2002–2007 Wissenschaftliche Mitarbeiter am Institut für Medizinische Biometrie und Epidemiologie (Prof. Dr. Schäfer), Philipps-Universität Marburg

1998–2001 Wissenschaftliche Hilfskraft im DFG-Projekt „Dynamik mentaler Repräsentationen“ (Prof. Dr. Rösler), Philipps-Universität Marburg

Studium und Weiterbildung

seit 2006 Masterstudiengang „Medical Biometry/Biostatistics“, Ruprecht-Karls-Universität Heidelberg

2003–2007 Promotionsstudium Ergänzungsstudiengang „Humanbiologie“, Philipps-Universität Marburg

2002–2007 Postgraduelle Ausbildung „Medizinische Biometrie“, Ruprecht-Karls-Universität Heidelberg,
(Abschlussnote: summa cum laude)

1996–2002 Studium der Psychologie, Philipps-Universität Marburg,
Diplomarbeit „Gibt es räumlich getrennte semantische Repräsentationen von Nomen und Verben im Mentalen Lexikon ?“ (Prof. Rösler), Diplom in Psychologie (Note: sehr gut)
1995–1996  Studium der Architektur, Bauhausuniversität Weimar,  
(Exmatrikulation auf eigenen Wunsch)

Auslandsaufenthalt
2001–2002  Forschungsaufenthalt am „Brain Development Lab“  
(Prof. Dr. Neville), University of Oregon Eugene, USA

Schulbildung und Wehrersatzdienst
1994–1995  Wehrersatzdienst im Bereich Altenpflege, Diakonisches Werk, Bad Hersfeld
1981–1994  Grundschule, Gymnasialzweig und gymnasiale Oberstufe (Allgemeine Hochschulreife), Bad Hersfeld
Eigene Publikationen


Verzeichnis der akademischen Lehrer

Meine akademischen Lehrer waren die Damen und Herren

in Eugene, Oregon, USA
Neville

in Heidelberg
Abel, Bender, Brunner, Büchler, Chang-Claude, Dahlhaus, Diepgen, Eckhart, Hauschke, Haefeli, Hothorn L, Hund, Kieser, Nobiling, Mammen, Martus, Meinck, Molenberghs, Mikus, Röhmel, Sickmüller, Verbeke, Victor, Wellek, Wildemann

in Leipzig
von Cramon

in Marburg
Florin, Grzeschik, Hebebrand, Jacobs, Kleinmann, Lachnit, Lenz, Lohaus, Müller, Remschmidt, Rief, Röder, Röhrle, Rösler, Rost, Schäfer, Scheiblechner, Schmidt-Atzert, Schulze, Schwarting, Sommer, Staufenbiel, Stelzl, Stemmler, Wagner
Danksagung

An dieser Stelle möchte ich verschiedenen Personen danken, die mich beim Erstellen dieser Arbeit unterstützt haben.


Als bereichernd und herausfordernd empfand ich die Kooperationen mit klinischen und molekulargenetischen Partnern. Hierbei ist die Zusammenarbeit mit Herrn Prof. Dr. Johannes Hebebrand und Dr. Anke Hinney und allen Marburger und Essener Mitarbeiterinnen und Mitarbeitern hervorzuheben. Neben der fachlichen Ebene trug diese Kooperation auch nicht unwesentlich zu meinem privaten Glück bei. Zusätzlich möchte ich die Kooperation mit den GEM Heidelberg (Prof. Dr. Jenny Chang-Claude, Dr. Lars
Beckmann und die anderen Kolleginnen und Kollegen) herausstellen; ich empfand unsere
lockerer Gespräche über wissenschaftliche Inhalte als sehr bereichernd.

Ich danke meinen Eltern und meiner Schwester dafür, dass sie mich stets ermutigt
haben und mein steter Rückhalt waren. Allen anderen Verwandten und Freunden danke
ich für die Ablenkungen von der Arbeit, gemeinsam verbrachte Stunden sowie Hilfe beim
Korrekturlesen (Keri, Nina, Roman, Susanne und Susann).

Schließlich gilt ein besonderer Dank meiner Freundin Susann Friedel für ihre fortwährende
Unterstützung.
Ehrenwörtliche Erklärung


Essen, den 18.10.2007
A Appendix for chapter 3

A.1 C++ program for family-based association analyses

```cpp
#include <cstdlib>
#include <iostream>
#include <fstream>
#include <string>
#include <math.h>
using namespace std;

// determine the number of markers that need to be evaluated
void size_of_fields(int& c_geno, int& individuals, int marker, string filename, int& error) {
    int counter=0;
    string emptyvalue;
    ifstream file_in;
    file_in.open(filename.c_str(),ios_base::in);
    if (!file_in){
        cerr<<"An error occurred!";
        error=1;
    } else{
        while (file_in) {
            file_in>>emptyvalue;
            counter++;
        }
    }
}
```

// Program for construction case/pseudocontrols according to
// for family-based association studies using conditional logistic
// regression
// - takes a standard linkage file as input
// - generates 3 pseudocontrols for each case which is written to an outputfile
// called "PCC_input.txt" (put in the folder from where the program is
called)
// - checks for genotyping errors/missing genotypes and Mendelian inheritance
// - allows to delete uninformative families from the output (Option u)
// - allows to preserve parent-of-origin information (Option p)
// - allows to procedure a r-script "r-script.r" to do
// conditional logistic regression analyses (Option r)

// Copyright Andre Scherag,
// Institute of Med. Biometry and Epidemiology
// Philipps-University, Marburg, Germany
// Version 1.0 (01/2006)
```
(file_in.close();
individuals=counter/(6+2*marker);
c_gen5=6+2*marker;
);

// reads the complete linkage file and fills the genotype array
void linkage_read(string filename,string** geno,int c_geno,int
individuals){

ifstream file_in;
file_in.open(filename.c_str(),ios_base::in);
if (!file_in){
cerr<<"An error occured!";
}
else{
for (int i = 0; i < individuals; i++){
for (int j = 0; j < c_geno; j++){
    file_in>>geno[i][j];
};
}
file_in.close();
};

// reads the parental genotype
void Parentalgenotype(string FamID,string Identifier,int counter,
int Personen,string** Geno,string& A1,string& A2){

A1="NA";
A2="NA";
for (int i = 0; i < Personen; i++){
    if ((Geno[i][0]==FamID) & (Geno[i][1]==Identifier)) {
        if ((Geno[i][counter])!="0") A1=Geno[i][counter];
        if ((Geno[i][counter+1])!="0") A2=Geno[i][counter+1];
    };
}

// determines the case/offspring genotypes
void Childgenotype(string A1, string A2,string PK1,string PK2,string
PK3,string PK4,string& Child){

if (A1+A2==PK1) Child=PK1;
if (A2+A1==PK1) Child=PK1;
if (A1+A2==PK2) Child=PK2;
if (A2+A1==PK2) Child=PK2;
if (A1+A2==PK3) Child=PK3;
if (A2+A1==PK3) Child=PK3;
if (A1+A2==PK4) Child=PK4;
if (A2+A1==PK4) Child=PK4;
if ((A1=="0")&(A2=="0")&((PK1==PK2)&(PK1==PK3)&(PK1==PK4)&(PK2==PK3)
    &((PK2==PK4)&(PK3==PK4))) Child="NA";
};

// generates case/pseudocontrol output array
void pseudocontrols(string** Geno, int S_Geno, int Personen,
        string* & output, int n_trios, int option, int option2)
{
    string FamID, PerID, Child1, Child2, Child, Childsex;
    string MotherA1, MotherA2, FatherA1, FatherA2, PK1, PK2, PK3, PK4, gmo, gfa;
    int GenotypingError = 0;
    int MendelianError = 0;
    int doublehet = 0;
    int trios = 0;
    int Marker = (S_Geno - 6) / 2;

    for (int i = 0; i < Personen; i++)
    {
        FamID = (Geno[i][0]);
        PerID = (Geno[i][1]);
        Childsex = (Geno[i][4]);

        // offspring detected
        if (((Geno[i][2]) != "0") & ((Geno[i][3]) != "0"))
        {
            output[(trios)][0] = FamID;
            output[(trios+1)][0] = FamID;
            output[(trios+2)][0] = FamID;
            output[(trios+3)][0] = FamID;
            output[(trios)][1] = Childsex;
            output[(trios+1)][1] = Childsex;
            output[(trios+2)][1] = Childsex;
            output[(trios+3)][1] = Childsex;
            output[(trios)][2] = "1";  // case/pseudocontrol status
            output[(trios+1)][2] = "0";
            output[(trios+2)][2] = "0";
            output[(trios+3)][2] = "0";
            output[(trios)][3] = "1";  // status for cond. log. regression
            output[(trios+1)][3] = "2";
            output[(trios+2)][3] = "2";
            output[(trios+3)][3] = "2";

            for (int j = 0; j < Marker; j++)
            {
                int k = 6 + j * 2;

                Child1 = (Geno[i][k]) + (Geno[i][k+1]);
                Child2 = (Geno[i][k+1]) + (Geno[i][k]);

                Parentalgenotype(FamID, (Geno[i][3]), k, Personen, Geno, MotherA1, MotherA2);
                Parentalgenotype(FamID, (Geno[i][2]), k, Personen, Geno, FatherA1, FatherA2);

                // genotyping error/missing genotype
                if (((Geno[i][k]) == "0") | (Geno[i][k+1] == "0") |
                    (MotherA1 == "NA") | (MotherA2 == "NA") |
                    (FatherA1 == "NA") | (FatherA2 == "NA"))
                {
                    GenotypingError++;
                    output[(trios)][j+4] = "NA";
                    output[(trios+1)][j+4] = "NA";
                    output[(trios+2)][j+4] = "NA";
                    output[(trios+3)][j+4] = "NA";
                }
                else
                {
                    PK1 = MotherA1 + FatherA1;
                    PK2 = MotherA1 + FatherA2;
                    PK3 = MotherA2 + FatherA1;
                    PK4 = MotherA2 + FatherA2;
                }
        }
    }
}
gmo=max(MotherA1,MotherA2)+min(MotherA1,MotherA2);
gfa=max(FatherA1,FatherA2)+min(FatherA1,FatherA2);

//Mendelian error
if ((Geno[i][k]!="0")&(Geno[i][k+1]!="0")&
(Child1!=PK1)&&(Child1!=PK2)&&(Child1!=PK3)&
(Child1!=PK4)&&(Child1!=PK2)&&(Child2!=PK1)&&(Child2!=PK4))
    
MendelianError++;
output[(trios)][j+4] = "NA";
output[(trios+1)][j+4]="NA";
output[(trios+2)][j+4]="NA";
output[(trios+3)][j+4]="NA";
else{

switch(option){
//neglect parental origin
    case 0:
        Child=max((Geno[i][k]),(Geno[i][k+1]))+
        min((Geno[i][k]),(Geno[i][k+1]));
        PK1=max(MotherA1,FatherA1)+min(MotherA1,FatherA1);
        PK2=max(MotherA2,FatherA1)+min(MotherA2,FatherA1);
        PK3=max(MotherA1,FatherA2)+min(MotherA1,FatherA2);
        PK4=max(MotherA2,FatherA2)+min(MotherA2,FatherA2);
        if (option2==0){
            output[trios][j+4]=Child;
            if (Child == PK1) {output[trios+1][j+4]=PK2;
                output[trios+2][j+4]=PK3;output[trios+3][j+4]=PK4;};
            if (Child == PK2) {output[trios][j+4]=PK1;
                output[trios+2][j+4]=PK3;output[trios+3][j+4]=PK4;};
            if (Child == PK3) {output[trios][j+4]=PK1;
                output[trios+2][j+4]=PK2;output[trios+3][j+4]=PK3;};
            if (Child == PK4) {output[trios][j+4]=PK1;
                output[trios+2][j+4]=PK2;output[trios+3][j+4]=PK3;};
        }
    

    else{
        //delete uninformative families
        if ((MotherA1==MotherA2)&(FatherA1==FatherA2)){
            output[(trios)][j+4] = "NA";
            output[(trios+1)][j+4]="NA";
            output[(trios+2)][j+4]="NA";
            output[(trios+3)][j+4]="NA";
        }
    
    else{
        output[trios][j+4]=Child;
        if (Child == PK1) {output[trios+1][j+4]=PK2;
            output[trios+2][j+4]=PK3;output[trios+3][j+4]=PK4;};
        if (Child == PK2) {output[trios+1][j+4]=PK1;
            output[trios+2][j+4]=PK3;output[trios+3][j+4]=PK4;};
        if (Child == PK3) {output[trios+1][j+4]=PK1;
            output[trios+2][j+4]=PK2;output[trios+3][j+4]=PK3;};
        
    }
}

//preserve parental origin
    case 1:
        if (option2==0){
            //heterozygous parents with same allele combination
            if ((MotherA1!=MotherA2) &

//child is homozygous
if ((Geno[i][k]==Geno[i][k+1]) &
        (Geno[i][k]!="0") &
        (Geno[i][k+1]!="0")){
    output[(trios)][j+4]=Geno[i][k]+Geno[i][k+1];
    if (MotherA1!=Geno[i][k])
        output[(trios+1)][j+4]=MotherA1+MotherA1;
    if (MotherA2!=Geno[i][k])
        output[(trios+2)][j+4]=MotherA2+MotherA2;
    output[(trios+3)][j+4]="NA";
}
//child is also heterozygous
else{
    doublehet++;
    output[(trios)][j+4]="NA";
    output[(trios+1)][j+4]="NA";
    output[(trios+2)][j+4]="NA";
    output[(trios+3)][j+4]="NA";
}
else{
    Childgenotype(Geno[i][k],Geno[i][k+1],PK1,PK2,PK3,PK4,Child);
    output[(trios)][j+4]=Child;
    if (Child == "NA"){
        GenotypingError++;
        output[(trios+1)][j+4]="NA";
        output[(trios+2)][j+4]="NA";
        output[(trios+3)][j+4]="NA";
    };
    if (Child == PK1){
        output[(trios+1)][j+4]=PK2;
        output[(trios+2)][j+4]=PK3;
        output[(trios+3)][j+4]=PK4;
    };
    if (Child == PK2){
        output[(trios+1)][j+4]=PK1;
        output[(trios+2)][j+4]=PK3;
        output[(trios+3)][j+4]=PK4;
    };
    if (Child == PK3){
        output[(trios+1)][j+4]=PK1;
        output[(trios+2)][j+4]=PK2;
        output[(trios+3)][j+4]=PK4;
    };
    if (Child == PK4){
        output[(trios+1)][j+4]=PK1;
        output[(trios+2)][j+4]=PK2;
        output[(trios+3)][j+4]=PK3;
    };
}
//delete uninformative families
else{
    if ((MotherA1==MotherA2) & (FatherA1==FatherA2)){
        output[(trios)][j+4]="NA";
        output[(trios+1)][j+4]="NA";
        output[(trios+2)][j+4]="NA";
        output[(trios+3)][j+4]="NA";
    };
}
//heterozygous parents with same allele combination
else{
    //heterozygous parents with same allele combination
    if ((MotherA1!=MotherA2) &
        (FatherA1!=FatherA2) &
        (gmo==gfa)){

}
if ((Geno[i][k] == Geno[i][k+1]) &
    (Geno[i][k] != "0") &
    (Geno[i][k+1] != "0")){
    output[(trios)][j+4]=Geno[i][k]+Geno[i][k+1];
    if (MotherA1 != Geno[i][k])
        output[(trios+1)][j+4]=MotherA1+MotherA1;
    if (MotherA2 != Geno[i][k])
        output[(trios+2)][j+4]=MotherA2+MotherA2;
    output[(trios+3)][j+4]="NA";
    output[(trios+4)][j+4]="NA";
}
//child is also heterozygous
else{
    doublehet++;
    output[(trios)][j+4] = "NA";
    output[(trios+1)][j+4] = "NA";
    output[(trios+2)][j+4] = "NA";
    output[(trios+3)][j+4] = "NA";
}
else{
    Childgenotype(Geno[i][k], Geno[i][k+1],
    PK1,PK2,PK3,PK4,Child);
    output[(trios)][j+4]=Child;
    if (Child == "NA"){
        GenotypingError++;
        output[(trios+1)][j+4]="NA";
        output[(trios+2)][j+4]="NA";
        output[(trios+3)][j+4]="NA";
    }
    if (Child == PK1){
        output[(trios+1)][j+4]=PK2;
        output[(trios+2)][j+4]=PK3;
        output[(trios+3)][j+4]=PK4;
    }
    if (Child == PK2){
        output[(trios+1)][j+4]=PK1;
        output[(trios+2)][j+4]=PK3;
        output[(trios+3)][j+4]=PK4;
    }
    if (Child == PK3){
        output[(trios+1)][j+4]=PK1;
        output[(trios+2)][j+4]=PK2;
        output[(trios+3)][j+4]=PK3;
    }
    if (Child == PK4){
        output[(trios+1)][j+4]=PK1;
        output[(trios+2)][j+4]=PK2;
        output[(trios+3)][j+4]=PK3;
    }
    }
};
break;
//end of switch
};
//end of Mendelian error
};
//end of genotyping error/missing genotype
};
//loop for each marker
trios=trios+4;
//loop for each child
};
//loop for each individual
n_trios=(trios/4);
cout <<"\n";
cout <<"P S E U D O C O N T R O L - C R E A T O R " <<"\n";
cout <<"****************************\n";
cout <<"\n";
cout <<"number of trios irresp. of family:\\t"<<n_trios<<"\n\n";
cout <<"number of genotyping errors/missing genotypes:\n";
cout <<"(# trios x # SNPs) :\\t"<<GenotypingError<<"\n";
double a1=GenotypingError;
double a2=n_trios*Marker;
double quote=a1/a2*100;
cout <<"\t in percent:\n"<ceil(quote)<<" %\n";
cout <<"number of Mendelian errors:\\t"<<MendelianError<<"\n";
quote=a3/a2*100;
cout <<"\t in percent:\n"<ceil(quote)<<" %\n";
if (option==1) cout <<"\nboth parents and child heterozygous:\n";
cout <<"\n";
cout <<"**************************************************************\n";
}

// write output from array to the PCC_input.txt
void file_write(int c_geno,int individuals,int n_trios,string** output){
    int num=1;
    int z=0;
    ofstream file_out;

    string filename="PCC_input.txt";
    file_out.open(filename.c_str(),ios_base::out);
    file_out<<"FAMNR ";    // variable names in headline
    file_out<<"FAMID ";
    file_out<<"CHILDSEX ";
    file_out<<"STATUS ";
    file_out<<"TIME ";

    for (int ik=1; ik<((c_geno-6)/2+1);ik++){
        file_out<<"SNP"<<ik<<" ";
    }
    file_out<<"\n";    //end of headline

    for (int ig=0; ig<(n_trios*4);ig++){
        file_out<<(num)<<" ";    //FAMNR
        z++;
        if (z==4) {
            num++;
            z=0;
        }
        for (int jg=0;jg<((((c_geno-6)/2+4)));jg++){
            file_out<< output[ig][jg]" ";
        }
        if (ig < ((n_trios*4)-1)) file_out<<"\n";
    }
    file_out.close();
}

// write R-script for the analysis
// -reads PCC_input.txt from R working directory
// -conducts cond. log. regression for each marker and
// saves genotype-wise z-Scores (res1) as well as
// global log likelihood differences (res2)
// -similarly analyses stratified by gender of child are included
void rscript(int marker,int n_trios,int individuals){

    // write output from array to the PCC_input.txt
    file_write(marker,n_trios,individuals,"PCC_input.txt");

    // read PCC_input.txt from R working directory
    // -conducts cond. log. regression for each marker and
    // saves genotype-wise z-Scores (res1) as well as
    // global log likelihood differences (res2)
    // -similarly analyses stratified by gender of child are included
}
ofstream file_out;
string filename="r-script.r";
file_out.open(filename.c_str(),ios_base::out);
file_out << "library(survival)\n";
file_out << "rawdata<-read.table("PCC_input.txt",header=TRUE)\n";
file_out << "femalechild<-rawdata[(rawdata[,3]=="2"),]\n";
file_out << "malechild<-rawdata[(rawdata[,3]=="1"),]\n";
file_out << "res1<-as.null(matrix(0,ncol=3,nrow=1))\n";
file_out << "res2<-as.null(matrix(0,ncol=1,nrow=1))\n";
file_out << "res1_f<-as.null(matrix(0,ncol=3,nrow=1))\n";
file_out << "res2_f<-as.null(matrix(0,ncol=1,nrow=1))\n";
file_out << "res1_m<-as.null(matrix(0,ncol=3,nrow=1))\n";
file_out << "res2_m<-as.null(matrix(0,ncol=1,nrow=1))\n";

for (int ik=1; ik<marker+1;ik++){
    if ((individuals/3)==n_trios) {
        file_out << "if (dim(table(rawdata$SNP" << ik << ")>)1) {\n";
        file_out << "y<-coxph(Surv(TIME,STATUS)~factor(SNP" << ik << ")+\n";
        file_out << "strata(FAMNR),rawdata,method="exact")\n";
        file_out << "if (dim(table(rawdata$SNP" << ik << ")==2 )\n";
        file_out << "res1<-rbind(res1,(diag(y$coef/sqrt(y$var)),NA,NA))}\n";
        file_out << "if (dim(table(rawdata$SNP" << ik << ")==3 )\n";
        file_out << "res1<-rbind(res1,(diag(y$coef/sqrt(y$var)),NA))}\n";
        file_out << "if (dim(table(rawdata$SNP" << ik << ")==4 )\n";
        file_out << "res1<-rbind(res1,diag(y$coef/sqrt(y$var)))}\n";
        file_out << "res2<-rbind(res2,(y$loglik[1]-y$loglik[2]))}\n";
        file_out << "} else {\n";
        file_out << "res1<-rbind(res1,c(1,1,1))\n";
        file_out << "res2<-rbind(res2,1)}\n";
    }
    else {
        file_out << "if (dim(table(rawdata$SNP" << ik << ")>)1) {\n";
        file_out << "y<-coxph(Surv(TIME,STATUS)~factor(SNP" << ik << ")+\n";
        file_out << "strata(FAMNR)+cluster(FAMID),robust=TRUE,rawdata)\n";
        file_out << "if (dim(table(rawdata$SNP" << ik << ")==2 )\n";
        file_out << "res1<-rbind(res1,(diag(y$coef/sqrt(y$var)),NA,NA))}\n";
        file_out << "if (dim(table(rawdata$SNP" << ik << ")==3 )\n";
        file_out << "res1<-rbind(res1,(diag(y$coef/sqrt(y$var)),NA))}\n";
        file_out << "if (dim(table(rawdata$SNP" << ik << ")==4 )\n";
        file_out << "res1<-rbind(res1,diag(y$coef/sqrt(y$var)))}\n";
        file_out << "res2<-rbind(res2,(y$loglik[1]-y$loglik[2]))}\n";
        file_out << "} else {\n";
        file_out << "res1<-rbind(res1,c(1,1,1))\n";
        file_out << "res2<-rbind(res2,1)}\n";
    }
}
else {
    if ((individuals/3)==n_trios) {
        file_out << "if (dim(table(femalechild$SNP" << ik << ")>)1) {\n";
        file_out << "y<-coxph(Surv(TIME,STATUS)~factor(SNP" << ik << ")+\n";
        file_out << "strata(FAMNR),femalechild,method="exact")\n";
        file_out << "if (dim(table(femalechild$SNP" << ik << ")==2 )\n";
        file_out << "res1_f<-rbind(res1_f,(diag(y$coef/sqrt(y$var)),NA,NA))}\n";
        file_out << "if (dim(table(femalechild$SNP" << ik << ")==3 )\n";
        file_out << "res1_f<-rbind(res1_f,(diag(y$coef/sqrt(y$var)),NA))}\n";
        file_out << "if (dim(table(femalechild$SNP" << ik << ")==4 )\n";
        file_out << "res1_f<-rbind(res1_f,diag(y$coef/sqrt(y$var)))}\n";
    }
    else {
        file_out << "if (dim(table(femalechild$SNP" << ik << ")>)1) {\n";
        file_out << "y<-coxph(Surv(TIME,STATUS)~factor(SNP" << ik << ")+\n";
        file_out << "strata(FAMNR),femalechild,method="exact")\n";
        file_out << "if (dim(table(femalechild$SNP" << ik << ")==2 )\n";
        file_out << "res1_f<-rbind(res1_f,(diag(y$coef/sqrt(y$var)),NA,NA))}\n";
        file_out << "if (dim(table(femalechild$SNP" << ik << ")==3 )\n";
        file_out << "res1_f<-rbind(res1_f,(diag(y$coef/sqrt(y$var)),NA))}\n";
        file_out << "if (dim(table(femalechild$SNP" << ik << ")==4 )\n";
        file_out << "res1_f<-rbind(res1_f,diag(y$coef/sqrt(y$var)))}\n";
    }
};
file_out<<"res2_f<-rbind(res2_f,(y$loglik[1]-y$loglik[2]))\"n;  
file_out<<"} else {\n;  
file_out<<"res1_f<-rbind(res1_f,c(1,1,1))\"n;  
file_out<<"res2_f<-rbind(res2_f,1)\"n;  
file_out<<"}\n;  
}  
else{  
file_out<<"if (dim(table(femalechild$SNP<<ik<<"))>1) {\n;  
file_out<<"y<-coxph(Surv(TIME,STATUS)~factor(SNP<<ik<<"),\n;  
file_out<<"strata(FAMNR)+cluster(FAMID),robust=TRUE,femalechild)\"n;  
file_out<<"if (dim(table(femalechild$SNP<<ik<<"))==2) {\n;  
file_out<<"res1_f<-rbind(res1_f,c(diag(y$coef/sqrt(y$var)),NA,NA))\"n;  
file_out<<"if (dim(table(femalechild$SNP<<ik<<"))==3) {\n;  
file_out<<"res1_f<-rbind(res1_f,c(diag(y$coef/sqrt(y$var)),NA))\"n;  
file_out<<"if (dim(table(femalechild$SNP<<ik<<"))==4) {\n;  
file_out<<"res1_f<-rbind(res1_f,diag(y$coef/sqrt(y$var)))\"n;  
file_out<<"res2_f<-rbind(res2_f,(y$loglik[1]-y$loglik[2]))\"n;  
file_out<<"} else {\n;  
file_out<<"res1_f<-rbind(res1_f,c(1,1,1))\"n;  
file_out<<"res2_f<-rbind(res2_f,1)\"n;  
file_out<<"}\n;  
};  
if ((individuals/3)==n_trios) {  
file_out<<"if (dim(table(malechild$SNP<<ik<<"))>1) {\n;  
file_out<<"y<-coxph(Surv(TIME,STATUS)~factor(SNP<<ik<<"),\n;  
file_out<<"strata(FAMNR),malechild,method="exact")\"n;  
file_out<<"if (dim(table(malechild$SNP<<ik<<"))==2) {\n;  
file_out<<"res1_m<-rbind(res1_m,c(diag(y$coef/sqrt(y$var)),NA,NA))\"n;  
file_out<<"if (dim(table(malechild$SNP<<ik<<"))==3) {\n;  
file_out<<"res1_m<-rbind(res1_m,c(diag(y$coef/sqrt(y$var)),NA))\"n;  
file_out<<"if (dim(table(malechild$SNP<<ik<<"))==4) {\n;  
file_out<<"res1_m<-rbind(res1_m,diag(y$coef/sqrt(y$var)))\"n;  
file_out<<"res2_m<-rbind(res2_m,(y$loglik[1]-y$loglik[2]))\"n;  
file_out<<"} else {\n;  
file_out<<"res1_m<-rbind(res1_m,c(1,1,1))\"n;  
file_out<<"res2_m<-rbind(res2_m,1)\"n;  
file_out<<"}\n;  
}  
else{  
file_out<<"if (dim(table(malechild$SNP<<ik<<"))>1) {\n;  
file_out<<"y<-coxph(Surv(TIME,STATUS)~factor(SNP<<ik<<"),\n;  
file_out<<"strata(FAMNR)+cluster(FAMID),robust=TRUE,malechild)\n;  
file_out<<"if (dim(table(malechild$SNP<<ik<<"))==2) {\n;  
file_out<<"res1_m<-rbind(res1_m,c(diag(y$coef/sqrt(y$var)),NA,NA))\n;  
file_out<<"if (dim(table(malechild$SNP<<ik<<"))==3) {\n;  
file_out<<"res1_m<-rbind(res1_m,c(diag(y$coef/sqrt(y$var)),NA))\n;  
file_out<<"if (dim(table(malechild$SNP<<ik<<"))==4) {\n;  
file_out<<"res1_m<-rbind(res1_m,diag(y$coef/sqrt(y$var)))\n;  
file_out<<"res2_m<-rbind(res2_m,(y$loglik[1]-y$loglik[2]))\n;  
file_out<<"} else {\n;  
file_out<<"res1_m<-rbind(res1_m,c(1,1,1))\n;  
file_out<<"res2_m<-rbind(res2_m,1)\n;  
file_out<<"}\n;  
};  
file_out.close();
int c_geno=0; // # columns of linkage (input) file
int individuals=0; // # individuals
int m=0; // # number of SNP markers
int n_trios=0; // # trios (irrespective of # of sibs/family)
string filename; // name of linkage (input) file
int error=0; // error check
int option=0;  // 0 (default)
    // 1 with preserved parental origin
int option2=0; // 0 (default)
    // 1 delete uninformative families
int option3=0; // 0 (default) no r-script;
    // 1 r-script will be produced

if (argc<2) {
    error=1;
    cout <<"\n"
    cout <<"**************************************************************\n";
    cout <<"P S E U D O C O N T R O L - C R E A T O R ********************\n";
    cout <<"**************************************************************\n";
    cout <<"this tool takes a linkage input file with nuclear families\n";
    cout <<"and affected offspring to create an output file with\n";
    cout <<"cases and pseudocontrols following the ideas of\n";
    cout <<"Cordell et al. (2004) and Cordell & Clayton (2002)\n";
    cout <<"\n";
    cout <<"CALL: PCC inputfile.txt (number of SNP markers) (Option)\n";
    cout <<"Options are:\n";
    cout <<"u (deletes uninformative families)\n";
    cout <<"p (preserve parental origin information)\n";
    cout <<"r (creates an r-script for analysis)\n";
    cout <<"**************************************************************\n";
};

if (error==0) {

    if (!((argv[1]||argv[2])){
        error=1;
    } else{
        filename=argv[1];
        n=atoi(argv[2]);
    }

    for (int y=3;y<argc;y++){
        if (strcmp(argv[y],"p")==0) option=1;
        if (strcmp(argv[y],"u")==0) option2=1;
        if (strcmp(argv[y],"r")==0) option3=1;
    };

    //determine size of arrays
    size_of_fields(c_geno,individuals,m,filename,error);

    if (m==0) error=1;
    if (error==0){

        //initialize field for genotype data
        string** geno=new string*[individuals];
        for (int mg=0; mg<individuals; mg++) geno[mg]=new string[(c_geno)];
        for (int ig=0; ig<individuals; ig++)
            for (int jg=0; jg<(c_geno); jg++) geno[ig][jg]="0";

        //read input linkage file and write it to array
        linkage_read(filename,geno,c_geno,individuals);

        //initialize output field
string** output=new string*[4*(individuals+4)];
for (int mg=0; mg<4*(individuals+4); mg++)
    output[mg]=new string[(((c_geno-6)/2+4))];
for (int ig=0; ig<4*(individuals+4); ig++)
    for (int jg=0; jg<(((c_geno-6)/2+4)); jg++) output[ig][jg]="0";

//produce output for conditional logistic regression
pseudocontrols(geno,c_geno,individuals,output,n_trios,option,option2);

//write output to a file
file_write(c_geno,individuals,n_trios,output);

//if wanted produce r script for automatized analysis
if (option3==1) rscript((c_geno-6)/2,n_trios,individuals);
B Appendix for chapter 4

B.1 C++ program for multi-stage candidate gene association studies

```
#include <cstdlib>
#include <iostream>
#include <stdio.h>
#include <math.h>
// use GNU scientific library packages - put .dll files in project folder
#include <gsl/gsl_integration.h>
#include <gsl/gsl_cdf.h>

#define pi 4.0 * atan(1.0)
#define constante 1.0 /sqrt(2*pi)

using namespace std;

// functions without comment are either used for
// numerical integration or standardization

double f(double x,void* params) {
    double f=constante*exp(-0.5*x*x);
    return f;
}

double NORMALP(double x,double m,double v){
    gsl_integration_workspace * w=gsl_integration_workspace_alloc(1000);
    double result,error;
    double alpha=1;
    gsl_function F;
    F.function=&f;
    F.params=&alpha;
    // Group-sequential plans for normally distributed outcomes:
    // - uses numerical integration based on Weddles rule
    // - early stopping with rejection on the null hypothesis
    // - alpha-sending with asymmetric critical values
    // Copyright Andr Scherag,
    // Institute of Med. Biometry and Epidemiology
    // Philipps-University, Marburg, Germany
    // Version 1.0 (10/2006)
    //-------------------------------------------------------------------------------
```
size_t tneval;
double gert=(x - m)/(sqrt(v));
gsl_integration_qag (&F,-gert,gert,1e-13,1e-13,1000,6,&result,&error);
return (result/2);
}

void STD(int n,double*& numerator,double denominator) {
    for (int i=0;i<(n+1);i++){
        numerator[i]=numerator[i]/(denominator);
    }
}

void STDEV(int n,double*& time) {
    for (int i=0;i<(n+1);i++){
        time[i]=sqrt(time[i]);
    }
}

double CONVOLUTION(int j,int points,double con,double factor,double xvalue,double* low_limit,double* upp_limit,double** mem){
    int i,jm;
    double prob, prob0, prob1, prob2, prob3, prob4, prob5, prob6;
    jm=j-1;
    if (jm==1){
        prob=con * exp(factor * xvalue * xvalue);
    } else{
        double bwidth=(fabs(low_limit[jm]-upp_limit[jm]))/(2*points);
        xvalue= xvalue - low_limit[jm];
        prob0 = 0.0; for (i = 0; i < (2*points); i += 6)
            prob0 = prob0 + mem[jm][i] * 
                exp(factor * (xvalue - i * bwidth)*(xvalue - i * bwidth));
        prob1 = 0.0; for (i = 1; i <= (2*points); i += 6)
            prob1 = prob1 + mem[jm][i] * 
                exp(factor * (xvalue - i * bwidth)*(xvalue - i * bwidth));
        prob2 = 0.0; for (i = 2; i <= (2*points); i += 6)
            prob2 = prob2 + mem[jm][i] * 
                exp(factor * (xvalue - i * bwidth)*(xvalue - i * bwidth));
        prob3 = 0.0; for (i = 3; i <= (2*points); i += 6)
            prob3 = prob3 + mem[jm][i] * 
                exp(factor * (xvalue - i * bwidth)*(xvalue - i * bwidth));
        prob4 = 0.0; for (i = 4; i <= (2*points); i += 6)
            prob4 = prob4 + mem[jm][i] * 
                exp(factor * (xvalue - i * bwidth)*(xvalue - i * bwidth));
        prob5 = 0.0; for (i = 5; i <= (2*points); i += 6)
            prob5 = prob5 + mem[jm][i] * 
                exp(factor * (xvalue - i * bwidth)*(xvalue - i * bwidth));
        prob6 = 0.0; for (i = 6; i <= (2*points); i += 6)
            prob6 = prob6 + mem[jm][i] * 

time[i]=sqrt(time[i]);
exp(factor * (xvalue - i * bwidth)*(xvalue - i * bwidth));
prob = bwidth * con * (41*prob0 + 216*prob1 + 27*prob2 + 272*prob3 +
27*prob4 + 216*prob5 + 41*prob6) / 140;
} return prob;
}

//-------------------------------------------------------------------------------

double PROBLOW(int j,int points,double* low_limit,double* plow,double* pupp,double** edge){
    int i;
    double prob,prob0,prob1,prob2,prob3,prob4,prob5,prob6;
    double awidth=(fabs(-15-low_limit[j]))/(2*points);
    prob0 = 0.0; for (i = 0; i < (2*points); i += 6) prob0= prob0+edge[j][i];
    prob1 = 0.0; for (i = 1; i <= (2*points); i += 6) prob1= prob1+edge[j][i];
    prob2 = 0.0; for (i = 2; i <= (2*points); i += 6) prob2= prob2+edge[j][i];
    prob3 = 0.0; for (i = 3; i <= (2*points); i += 6) prob3= prob3+edge[j][i];
    prob4 = 0.0; for (i = 4; i <= (2*points); i += 6) prob4= prob4+edge[j][i];
    prob5 = 0.0; for (i = 5; i <= (2*points); i += 6) prob5= prob5+edge[j][i];
    prob6 = 0.0; for (i = 6; i <= (2*points); i += 6) prob6= prob6+edge[j][i];
    prob = awidth * (41*prob0 + 216*prob1 + 27*prob2 + 272*prob3 +
27*prob4 + 216*prob5 + 41*prob6) / 140;
    return prob;
}

//-------------------------------------------------------------------------------

double PROBUPP(int j,int points,double* upp_limit,double* plow,double* pupp,double** field){
    int i;
    double prob,prob0,prob1,prob2,prob3,prob4,prob5,prob6;
    double awidth=(fabs(low_limit[j]-upp_limit[j]))/(2*points);
    prob0 = 0.0; for (i = 0; i < (2*points); i += 6) prob0= prob0+field[j][i];
    prob1 = 0.0; for (i = 1; i <= (2*points); i += 6) prob1= prob1+field[j][i];
    prob2 = 0.0; for (i = 2; i <= (2*points); i += 6) prob2= prob2+field[j][i];
    prob3 = 0.0; for (i = 3; i <= (2*points); i += 6) prob3= prob3+field[j][i];
    prob4 = 0.0; for (i = 4; i <= (2*points); i += 6) prob4= prob4+field[j][i];
    prob5 = 0.0; for (i = 5; i <= (2*points); i += 6) prob5= prob5+field[j][i];
    prob6 = 0.0; for (i = 6; i <= (2*points); i += 6) prob6= prob6+field[j][i];
    prob = plow[j] + awidth * (41*prob0 + 216*prob1 + 27*prob2 + 272*prob3 +
27*prob4 + 216*prob5 + 41*prob6) / 140;
    return prob;
}

//-------------------------------------------------------------------------------

void REST(int n,int j,int points,double con,double factor,double* low_limit,double* upp_limit,double** field,double**& edge){
    int steps;
    double awidth=(fabs(-15-low_limit[j]))/(2*points);
    for (steps = 0; steps <= (2*points); steps++){
        edge[j][steps] = CONVOLUTION(j,points,con,factor,(-15+steps*awidth),
        low_limit,upp_limit,field);
    }
}

//-------------------------------------------------------------------------------

void HILL(int n,int j,int points,double con,double factor,
double* low_limit,double* upp_limit,double**& field){
    int steps;
}
double awidth=(fabs(low_limit[j]-upp_limit[j]))/(2*points);
for (steps = 0; steps <= (2*points); steps++)
    field[j][steps] = CONVOLUTION(j,points,con,factor,
      (low_limit[j]+steps*awidth),low_limit,upp_limit,field);
}

// calculate the upper and lower rejection errors
void SEQ(int n,int points,double* tscale,double* low_limit,double* upp_limit,
  double*& plow, double*& pupp){

double prob,awidth, uwidth,width,factor,con;
double** edge=new double*[n+1];
for (int i=0; i<(n+1); i++)
    edge[i]=new double[(2*points)+1];
double** field=new double*[n+1];
for (int i=0; i<(n+1); i++)
    field[i]=new double[(2*points)+1];

for (int j = 1; j < (n+1); j++)
    if (low_limit[j]<upp_limit[j]){
        factor = -.5 / (tscale[j]);
        con = constante / sqrt(tscale[j]);

        if (low_limit[j]<-50) {low_limit[j]=-50;};
        if (upp_limit[j]>50) {upp_limit[j]=50;};
        if (low_limit[j]>50) {low_limit[j]=50;};
        if (upp_limit[j]<-50) {upp_limit[j]=-50;};

        REST(n,j,points,con,factor,low_limit,upp_limit,field, edge);
        HILL(n,j,points,con,factor,low_limit,upp_limit,field);
        plow[j] = PROBLOW(j,points,low_limit,plow,pupp,edge);
        pupp[j] = PROBUPP(j,points,upp_limit,low_limit,plow,pupp,field);
    } else{
        printf("ERROR!\n");
    }
}

for (int i=0; i<(n+1); i++) delete[] edge[i];
delete[]edge;
for (int i=0; i<n+1; i++) delete[] field[i];
delete[]field;
}

// determine drift of Brownian motion that leads to specified power
void SEQDRIFT(int n,int points,double epsilon,double beta,double* times,double* b_low,double* b_upp,double& deltastar,double*& delta_low,double*& delta_upp){

double dl,du,bet,delta;

}
double* tsc=new double[n+1];
double* low=new double[n+1];
double* upp=new double[n+1];
double* plow=new double[n+1];
double* pupp=new double[n+1];

for (int i=2; i<(n+1); i++){
    plow[1]=0;
pupp[1]=0;
tsc[i]=times[i]-times[i-1];
    plow[i]=0;
pupp[i]=0;
}
dl=0;
du=b_upp[n]+NORMALV((1-beta),0,1)*sqrt(times[n]);

while (fabs(bet-beta) > epsilon){
    delta=((du+dl)/2);
    for (int i=1; i<(n+1); i++){
        low[i]=b_low[i]-delta*times[i];
        upp[i]=b_upp[i]-delta*times[i];
    }
}

SEQ(n,points,tsc,low,upp,plow,pupp);
bet=0;
bet=pupp[n]-plow[n];
if (bet>beta){
    dl=delta;
}
else{
    du=delta;
}
for (int i=1; i<(n+1); i++){
    delta_low[i]=plow[i];
    delta_upp[i]=pupp[i];
}
deltastar=delta;
delete tsc;
delete low;
delete upp;
delete plow;
delete pupp;
}

// determine critical values for each stage
void MAKELIMIT(int n,int points,double epsilon,double* times,
    double* alpha_low,double* alpha_upp,
    double*& b_low,double*& b_upp){
    double loop,blu,bbu,bll,size_low,size_upp;
    size_low=0;
    size_upp=0;
    for (int j=1; j<(n+1); j++){
\[ b_{upp}[j] = \text{NORMALV}\left((1 - \alpha_{upp}[j]), 0, 1\right) \times \sqrt{\text{times}[j]} \]
\[ b_{low}[j] = \text{NORMALV}\left((\alpha_{low}[j]), 0, 1\right) \times \sqrt{\text{times}[j]} \]

for (int k=2; k<(n+1); k++) {
    double* tsc = new double[k+1];
    double* low = new double[k+1];
    double* upp = new double[k+1];
    double* plow = new double[k+1];
    double* pupp = new double[k+1];
    for (int i=2; i<(k+1); i++) {
        plow[1] = 0;
        pupp[1] = 0;
        tsc[1] = 1;
        tsc[i] = times[i] - times[i-1];
        plow[i] = 0;
        pupp[i] = 0;
    }
    blu = \text{NORMALV}\left((\alpha_{low}[k] + \alpha_{upp}[k]), 0, 1\right) \times \sqrt{\text{times}[k]};
    bll = \text{NORMALV}\left((\alpha_{low}[k] - \alpha_{low}[k-1]), 0, 1\right) \times \sqrt{\text{times}[k]};
    while (fabs(size_low - \alpha_{low}[k]) > \epsilon) {
        b_low[k] = (blu + bll) / 2;
        for (int i=1; i<(k+1); i++) {
            low[i] = b_low[i];
            upp[i] = b_upp[i];
        }
        SEQ(k, points, tsc, low, upp, plow, pupp);
        size_low = 0;
        for (int j=1; j<(k+1); j++) {
            size_low = size_low + plow[j];
        }
        if (size_low > \alpha_{low}[k]) {
            blu = b_low[k];
        } else {
            bll = b_low[k];
        }
    }
    b_low[k] = blu;
    for (int i=2; i<(k+1); i++) {
        plow[1] = 0;
        pupp[1] = 0;
        plow[i] = 0;
        pupp[i] = 0;
    }
    buu = \text{NORMALV}\left(1 - (\alpha_{upp}[k] - \alpha_{upp}[k-1]), 0, 1\right) \times \sqrt{\text{times}[k]};
    bul = b_upp[k];
    while (fabs(size_upp - \alpha_{upp}[k]) > \epsilon) {
        b_upp[k] = (bul + buu) / 2;
        for (int i=1; i<(k+1); i++) {
            low[i] = b_low[i];
            upp[i] = b_upp[i];
        }
        SEQ(k, points, tsc, low, upp, plow, pupp);
    }
    loop = 0;
}
for (int j=1; j<k; j++) {
    loop=loop+(plow[j]);
};

size_upp=1-pupp[k]-loop;
if (size_upp>alpha_upp[k]){
    bu=b_upp[k];
} else{
    buu=b_upp[k];
};

b_upp[k]=buu;
delete tsc;
delete low;
delete upp;
delete plow;
delete pupp;
};

//-----------------------------------------------------------------------------
// determine sample sizes for each stage
void SEQSAMPLE(int n,int points,double imue,double isigma,double 
    epsilon,double beta,double* times,double* b_low,double* b_upp, 
    double& deltastar,double fs){

double dl,du,bet,deltasample;

dl=0;

du=1000000;

bet=0;

deltasample=0;

double* tsc =new double[n+1];

double* low =new double[n+1];

double* upp =new double[n+1];

double* plow=new double[n+1];

double* pupp=new double[n+1];

for (int i=2; i<(n+1); i++){
    plow[i]=0;
    pupp[i]=0;
    tsc[i]=i;
    tsc[i]=times[i]-times[i-1];
    plow[i]=0;
    pupp[i]=0;
};

dl=0;

du=1000000;

while (fabs(bet-beta) > epsilon){
    deltasample=((du+dl)/2);
    for (int i=1; i<(n+1); i++){
        double helk=(sqrt(deltasample)*imue*times[i]);
        low[i]=(b_low[i]-helk)/isigma;
        upp[i]=(b_upp[i]-helk)/isigma;
    };
    SEQ(n,points,tsc,low,upp,plow,pupp);
    bet=0;
    bet=pupp[n]-plow[n];

    if (bet>beta){
        dl=deltasample;
    } else{
du=deltasample;
}
);
deltastar=deltasample;
delete tsc;
delete low;
delete upp;
delete plow;
delete pupp;
}

//ADDITIONAL FUNCTIONS ----------------------------------------

// calculate mue for TDT where dd is MAF, rr is Psi1, and ss is Psi2
double calc_mue(double dd,double rr,double ss){
    double calc_mue=(-1)*((sqrt(2)*(1-rr+dd*(-1+2*rr- ss))*
                                 sqrt(1+rr+dd*(-1+ss)))/(1+rr+dd*(-1+ss)));
    return calc_mue;
}

// calculate variance sigmasquare for TDT where dd is MAF, rr is
// Psi1, and ss is Psi2
double calc_sigmaq(double dd,double rr,double ss){
    if (dd==0.5) dd=dd+0.0000000000001;
    double s1= dd-((4*(-1+dd)*(-1+2*dd)*(1+dd*(-1+rr))*(1+dd*(-1+rr))*(1+rr)) /
                      (pow((1+rr+dd*(-1+ ss)),3)));
    double s2=((4*(1+dd*(-1+rr))*(-3-rr-3*dd*rr+dd*dd*(3+rr)))/
                   (pow((1+rr+dd*(-1+ ss)),2)));
    double s3=((-15+rr-dd*(-47+13*rr+4*dd*(7+dd*(-7*dd)*rr)))/
                      ((-1+2*dd)*(1+rr+dd*(-1+ ss))));
    double s4=(( 2*(-1+dd)*(1+dd*(-1+rr)))/
                      ((-1+2*dd)*(-1-dd)*(-1+2*rr)-dd+dd*ss)));
    double calc_sigmaq=0.25*(s1+s2+s3-s4);
    return calc_sigmaq;
}

// determine FS design sample size
double sample(double zalpha,double zpower,double mue,double sigma){
    double sample;
    if (mue!=0){
        sample=((zalpha+zpower*sigma)*(zalpha+zpower*sigma))/(mue*mue);
    }
    return sample;
}

//MAIN PROGRAM (HAS TO BE MODIFIED FOR ALTERNATIVE CALCULATIONS)

int main(int argc,char *argv[]){
//starts with a common GS design not --------
int n=3; //MODIFY: # of analysis

double* a_upp=new double[n+1]; //MODIFY: Cumulated upper alphas
    a_upp[1]=0.005;
a_upp[2]=0.015;
a_upp[3]=0.025;
a_upp[4]=0.025;

double* delta_upp=new double[n+1];
double* dprob_upp=new double[n+2];
double* expt_upp=new double[n+2];

double* a_low=new double[n+1]; //MODIFY: Cumulated lower alphas
    a_low[1]=0.0025;
a_low[2]=0.0075;
a_low[3]=0.0125;
a_low[4]=0.025;

double* delta_low=new double[n+1];
double* dprob_low=new double[n+2];
double* expt_low=new double[n+2];

double* times=new double[n+1]; //MODIFY: Information time
    times[0]=0;
times[1]=0.3;
times[2]=0.6;
times[3]=1;
    times[4]=1.0;

for (int j=0; j<(n+1); j++)
    ti[j]=times[j];
ti[n+1]=1;

double* b_upp=new double[n+1]; //critical values
double* b_low=new double[n+1];
double* nb_upp=new double[n+1];
double* nb_low=new double[n+1];

double beta=0.2; //MODIFY: Beta

double power=1-beta; //MODIFY: Power
    // also used for FS design

double epsilon=1.e-11; //precision for bi-sectional search

double deltastar=0;
double tupp=0;
double tlow=0;

int points; //MODIFY: # of sampling points (*6*2)
    //according to Weddles rule

points=6*10;

STD(n, times,(times[n])); //std all information times
STD(n, ti,(ti[n]));
STD(n, times,(times[1]));

for (int j=1; j<(n+1); j++) {
    // calculate limits
    b_upp[j] = NORMAL((1-a_upp[j]),0,1)*sqrt(times[j]);
    b_low[j] = NORMAL((a_low[j]),0,1)*sqrt(times[j]);
}

MAKELIMIT(n,points,epsilon,times,a_low,a_upp,b_low,b_upp);

SEQDRIFT(n,points,epsilon,beta,times,b_low,b_upp,
    deltagr,del`alp,del`upp);
    double del`astar_upp=deltastar;

    for (int j=2; j<(n+1); j++){
        //dprob_upp[j] = (delta_upp[j-1]-delta_low[j-1])-
        //delta_upp[j]*delta_low[j];
        dprob_upp[1] = 1-delta_upp[1]-delta_low[1];
        dprob_upp[j] = (delta_upp[j-1]-delta_low[j-1])-
            delta_upp[j]*delta_low[j];
        dprob_upp[n+1] = delta_upp[n]-delta_low[n];
        expt_upp[1] = dprob_upp[1]*ti[1];
        expt_upp[j] = dprob_upp[j]*ti[j];
        expt_upp[n+1] = dprob_upp[n+1]*ti[n+1];
    }

for (int j=1; j<(n+1); j++){
    nb_upp[j] = -1*b_low[j];
    nb_low[j] = -1*b_upp[j];
}

SEQDRIFT(n,points,epsilon,beta,times,nb_low,nb_upp,deltastar,
    del`alp,del`upp);
    double del`astar_low=deltastar;

    for (int j=2; j<(n+1); j++){
        //dprob_low[j] = (delta_upp[j-1]-delta_low[j-1])-
        //delta_upp[j]+delta_low[j];
        dprob_low[1] = 1-delta_upp[1]-delta_low[1];
        dprob_low[j] = (delta_upp[j-1]-delta_low[j-1])-
            delta_upp[j]+delta_low[j];
        dprob_low[n+1] = delta_upp[n]-delta_low[n];
        expt_low[1] = dprob_low[1]*ti[1];
        expt_low[j] = dprob_low[j]*ti[j];
        expt_low[n+1] = dprob_low[n+1]*ti[n+1];
    }

    for (int j=1; j<(n+2); j++){
        tupp=tupp+expt_upp[j];
        tlow=tlow+expt_low[j];
    }

printf("--------------------\n"); // display results
printf("Common GB design\n");
printf("--------------------\n\n");

printf("UPPER\n");

printf("nominal alphas\n");

for (int j=1; j<(n+1); j++){
    printf("stage %i\n",j);
    a_upp[j] = 0.5-NORMAL((b_upp[j]/(sqrt(times[j]))),0,1);
    printf("\t %.18f\n",a_upp[j]);
}

printf("critical values\n");

for (int j=1; j<(n+1); j++){
    printf("stage %i\n",j);
    printf("\t %.18f\n",(b_upp[j]/(sqrt(times[n]))));
}
printf("BM-drift\t\t % .18f\n\n",(deltastar_upp/sqrt(times[1]/times[n])));

printf("prob. of crossing the upper boundary exactly\n");
printf("at the respective time point (last value\n");
printf("is prob. of not crossing any boundary)\n")

for (int j=1; j<(n+2); j++){
    if (j<n+1) printf("stage % i",j);
    else printf("beta\t");
    printf("\t % .18f\n",(dprob_upp[j]));
}

printf("\naverage stopping time =\t % .18f\n",tupp);
printf("\n---------------------\n\nLOWER\n\n\nnominal alphas\n")

for (int j=1; j<(n+1); j++){
    printf("stage % i",j);
    a_low[j]= 0.5+NORMALP((b_low[j]/(sqrt(times[j]))),0,1);
    printf("\t % .18f\n",a_low[j]);
}

printf("\ncritical values\")

for (int j=1; j<(n+1); j++){
    printf("stage % i",j);
    printf("\t % .18f\n",(b_low[j]/(sqrt(times[n]))));
}

printf("BM-drift\t\t % .18f\n\n",(-deltastar_low/sqrt(times[1]/times[n])));

printf("prob. of crossing the lower boundary exactly\n");
printf("at the respective time point (last value\n");
printf("is prob. of not crossing any boundary)\n")

for (int j=1; j<(n+2); j++){
    if (j<n+1) printf("stage % i",j);
    else printf("beta\t");
    printf("\t % .18f\n",(dprob_low[j]));
}

printf("\naverage stopping time =\t % .18f\n",tlow);
printf("\n---------------------\n\n\n\nДефо: TDT - enter parameters of genetic model: GRRs and MAF

//NOTE: For a GS design changes of the alpha-spending sequence
// may be necessary!

double MAF=0.1; //MODIFY: TDT - enter parameters of
            // genetic model: GRRs and MAF

double psi1=1.5;
double psi2=2.25;

imue = calc_mue(MAF,psi1,psi2);
double isigmaq; // sigmasquare for TDT (see chapter 5)
isigmaq = calc_sigmaq(MAF, psi1, psi2);
// NOTE: imue and isigmaq can be changed and entered manually if no TDT design
// is foreseen (a R package by R.Pahl can be used to determine the values for
// the CA trend test)

printf("---------------------\n\n\nFS and GS designs

sample size calculations for a candidate gene association study\n\n\n---------------------\n\n\ndouble fs=0; // sample size - FS design
double gs=0; // sample size - GS design

if ((imue!=0)&&(isigmaq>0)){
    fs=ceil(sample((NORMALV((1-alpha/2),0,1)),(NORMALV((power),0,1)),
              imue,(sqrt(isigmaq))));
};
printf("total sample size FS design % .0f\n",fs);

double nstart=0;
SEQSAMPLE(n, points, imue, (sqrt(isigmaq)), epsilon, (1-power),
times, b_low, b_upp, nstart, fs);
gs=ceil(nstart*times[n]);
printf("total sample size GS design % .0f\n",gs);
printf("exp. sample size GS design (low) % .0f\n",ceil(gs*tlow));
printf("exp. sample size GS design (upp) % .0f\n",ceil(gs*tupp));

for (int j=1; j<(n+1); j++){
    gs=nstart*times[j];
    printf(" stage %i",j);
    printf(" cumulative sample size % .0f\n",ceil(gs));
};
delete a_upp; // delete arrays
delete a_low;
delete times;
delete b_upp;
delete b_low;
delete nb_upp;
delete nb_low;
system("PAUSE");
return EXIT_SUCCESS;"